Feed the Future

Survey Implementation

Document

Protocol for Preparing Non-Public, Restricted, and Public Access Datasets

Zone of Influence Survey

February 2020

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

**Recommended Citation**

Anna Brenes and Kiersten B. Johnson. 2020. Feed the Future Zone of Influence Surveys: Protocol for Preparing Non-Public, Restricted, and Public Access Datasets. Washington, DC: Bureau for Food Security, USAID.

**Contact Information**

Feed the Future

1300 Pennsylvania Ave, NW

Washington, DC 20004

[www.feedthefuture.gov](about:blank)

**TABLE OF CONTENTS**

[I. INTRODUCTION 3](#_Toc34734401)

[II. PREPARING DATA FOR REGISTRATION AND SUBMISSION TO THE USAID DEVELOPMENT DATA LIBRARY (DDL) 3](#_Toc34734402)

[III. INSTRUCTIONS FOR PREPARING NON-PUBLIC ACCESS DATASETS 5](#_Toc34734403)

[IV. INSTRUCTIONS FOR PREPARING RESTRICTED ACCESS DATASETS 6](#_Toc34734404)

[V. INSTRUCTIONS FOR PREPARING PUBLIC ACCESS DATASETS 7](#_Toc34734405)

[VI. REGISTERING AND SUBMITTING DATA TO THE DEVELOPMENT DATA LIBRARY (DDL) 13](#_Toc34734406)

[APPENDIX A: CONTACTS AND RESOURCES 19](#_Toc34734407)

[APPENDIX B: EXAMPLE PRIVACY ASSURANCE STATEMENT 20](#_Toc34734408)

[APPENDIX C: EXAMPLE DISCLOSURE ANALYSIS PLAN 21](#_Toc34734409)

[APPENDIX D: EXAMPLE AGRICULTURE MODULE TABLE FOR PUBLIC ACCESS DATASETS 28](#_Toc34734410)

# INTRODUCTION

Under the terms of USAID's [Development Data Policy](https://www.usaid.gov/ads/policy/500/579) and [Public Access Plan](https://www.usaid.gov/open/public-access-plan), USAID is making data from Feed the Future’s Zone of Influence Surveys available as open and machine-readable public use datasets available on USAID’s [Development Data Library (DDL)](https://data.usaid.gov/).

An essential step in the preparation of datasets for public use is ensuring respondent privacy and confidentiality. Respondent privacy and confidentiality is ensured and protected by removing and/or masking identifying information from the datasets, including *direct identifiers* (information such as names, addresses, Global Positioning System (GPS) coordinates, or any other personally identifying number or characteristic) and *indirect identifiers* (data that do not specifically identify a person or location, but that can be used to do so, one variable at a time or in combination, by uniquely describing a person or household).

The following guidance provides instructions for preparing non-public, restricted, and public use datasets for submission to USAID’s DDL and creating documentation that provides the user with the information essential to using the data. The documentation created following the guidance below must be included with submission of data by USAID survey implementing organizations to the DDL.

# II. PREPARING DATA FOR REGISTRATION AND SUBMISSION TO THE USAID DEVELOPMENT DATA LIBRARY (DDL)

USAID’s Bureau for Food Security requires partners who have implemented a Feed the Future Zone of Influence Survey to submit three types of datasets to the DDL:

* a Non-Public Access dataset that retains some direct and many of the high-risk indirect identifiers,
* a Restricted Access dataset that contains plot level geocoded data, and other more sensitive indirect identifiers necessary for internal and external analysis, and
* a Public Access dataset that has been processed to remove all direct identifiers and suppress or remove high-risk indirect identifiers.

This section provides instructions on how to prepare the Survey Description and ensure compliance with data use restrictions as a first step in the submission of Non-Public, Restricted, and Public Access datasets to the DDL.

**2.1 Preparing the Survey Description**

Provide a detailed description of the survey for all datasets. This information can be used to copy and paste to the DDL registration form Data Asset metadata tabs described in greater detail in the ***Registering and Submitting Data to the Development Data Library*** section of this document*.* Include the following information:

* A detailed description and extent of the geographic areas targeted by Feed the Future interventions, known as Feed the Future Zones of Influence (ZOI)
* Names of contractor(s) and subcontractor(s)
* Date range of when fieldwork took place
* Total number of households interviewed for the survey

EXAMPLE: In (COUNTRY), the (NAME OF SURVEY [e.g. 2019 Feed the Future Uganda Zone of Influence Survey]) fieldwork was conducted by (SUBCONTRACTOR), with technical assistance from (contractor). The fieldwork took place from (DATE - DD/MM/YYYY) and (DATE - DD/MM/YYYY). The ZOI in (COUNTRY) at the time of the survey covered (XX NUMBER) districts in (XX NUMBER) regions [provide detailed description of geographic area coverage of data collection including names of districts, regions or other administrative units]. A total of (XX NUMBER) enumeration areas were selected from the sampling frame, from which (XXXX NUMBER) households were interviewed for the survey.

**2.2 Ensuring Compliance with Data Use Restrictions**

Review your award (contract or grant) for relevant language on data use. All agreements with USAID/BFS or USAID missions, Institutional Review Boards (IRBs), as well as the survey questionnaire and informed consent statements must be reviewed to ensure that datasets at all levels of access are in compliance with those agreements and statements.

Approval to conduct the ZOI Survey in the country must have been obtained and documented. The approval letter from the host government and/or other regulatory body must clearly state that the implementing partner conducting the survey has full clearance to collect and disseminate the data following ethical data sharing practices.

The implementing partner’s IRB (or equivalent entity) should have provided clarification of the rights of children and a clear understanding of how their rights will be protected under local laws and ethical practices. Parental consent for the participation of children under age 18 years must have been obtained. If applicable, all restrictions on the collection and dissemination of the survey data should be documented as part of the Privacy Assurance Statement (see example in Appendix B).

**NOTE:** If it is determined that the host government will restrict data use, consult with the project AOR/COR, AO/CO, and USAID General Counsel/Regional Legal Officer/Authority.

# III. INSTRUCTIONS FOR PREPARING NON-PUBLIC ACCESS DATASETS

**Non-public access** files are for USAID archival and use purposes. These access files fall under “principled exceptions to the presumption in favor of openness” established in OMB Bulletin 12-01, “Guidance on Collection of U.S. Foreign Assistance Data.” For the Feed the Future Zone of Influence Surveys, the applicable exception is to the requirement that data that reveal private information about individuals must be kept confidential consistent with ethical guidelines and federal regulations (see ADS 508, Privacy Program). The exception applies to the household and plot location data (latitude and longitude), which are considered direct identifiers and could reveal private information about individuals. Non-public use data are only available for internal use by the Federal Government, such as by a single program, single agency, or across multiple agencies.

**3.1 Geographic detail**

Include the following:

* Household latitude and longitude coordinates: These data are essential for USAID to use in follow-on survey preparation and to inform evidenced-based analyses.
* Agricultural plot boundaries with attributes and geographic details (vertices, lines, polygon geometry)

**3.2 Address direct identifiers**

Remove the following:

* Names

**3.3 Create supporting documentation**

* Codebook: Include the original codebook
* Questionnaire: Include the final questionnaires used in the field (all languages)
* Copy of Informed Consent form(s)/statement(s) used: This can be uploaded to the DDL once for Non-Public, Restricted and Public Access datasets associated with a single Data Asset.
* Other supporting documentation: Include a brief README file providing details on the restriction of data use (internal only) and justification for why USAID needs the data for internal use.

**3.4 Label files appropriately**

Use the following naming convention (or similar) for all Non-Public Access dataset files to differentiate between Public, Restricted and Non-Public dataset files associated with a single Data Asset in USAID’s Development Data Library (DDL):

* EXAMPLE file name: FTF \_ZOI\_2019\_Kenya\_ Household\_NonPublic.csv. See Section VI below for more information on registering and submitting data to the DDL.

# IV. INSTRUCTIONS FOR PREPARING RESTRICTED ACCESS DATASETS

**Restricted Access** datasets may be shared with users who have been granted access with the necessary confidentiality pledge/non-disclosure agreement between the user organization and USAID. Restricted access files require that documentation of the syntax used to prepare and create intermediate and final computed variables and indicators and methodology for re-coding and suppression of variables be included with the data submission so that internal statistical tests of differences over time and other analyses can be done.

The following are steps required and a list of required documents to prepare Restricted Access datasets and upload them to the DDL.

**4.1 Mitigate geographic data risk**

Restricted Access datasets contain the full geographic detail of the plot level data collected, i.e., the exact boundary delineation of the agriculture plot. The household latitude and longitude data must be randomly offset following the guidance under Section V.

**4.2 Address direct and indirect identifiers**

The following direct identifiers should be offset, removed from, or top-coded in the Restricted Access data:

* + Names
  + Addresses
  + Any unique identifying number or characteristic
  + All elements of dates (except year) that are directly related to an individual, including birth date (NOTE: include year, month and day of birth for children < 6 years old, and year only for all other respondents)
  + Ages > 89 years: the upper end of the age distribution should be top-coded appropriately; see the section on data value re-coding, below.
  + Household latitude and longitude data must be offset following the guidance under Section V, *Instructions for Preparing Public Access Datasets*.

Include the following indirect identifiers:

* Plot boundaries and attributes (include original geographic detail)

**4.3 Create supporting documentation**

* Codebook: Include the original Codebook and the syntax used to recode, suppress and/or remove the original data.
* Questionnaire: Include the final questionnaires used in the field (all languages)
* Copy of Informed Consent form(s)/statement(s) used: This can be uploaded to the DDL once for Non-Public, Restricted and Public Access datasets associated with a single Data Asset.
* Other supporting documentation: Documentation should include information on how the data was collected, data use limitations, and any other contextual information that facilitates use of the Restricted Access data by other data analysts.

**4.4 Label files appropriately**

Use the following naming convention (or similar) for all Restricted Access dataset files to differentiate between Public, Restricted and Non-Public dataset files associated with a single Data Asset in USAID’s Development Data Library (DDL):

**EXAMPLE:** FTF \_ZOI\_2019\_Kenya\_ Household\_Restricted.csv

See Section VI below for more information on registering and submitting data to the DDL.

# V. INSTRUCTIONS FOR PREPARING PUBLIC ACCESS DATASETS

**Public Access Datasets** are or could be made publicly available to all without restrictions. All direct identifiers collected during the survey must be removed or sufficiently anonymized to protect the privacy of respondents. The following guidance provides details on mitigating and managing risk to ensure data can be made public.

**5.1 Mitigate geographic data risk**

If multiple variables collected in the survey identify the granular geographic location of respondents, the locations must be sufficiently anonymized so that individual households and/or respondents cannot be identified. These variables are necessary during data collection and data processing, and are needed for internal analysis, but they present a substantial risk to the anonymity of respondents. The following guidance provides instructions for preparing high-risk geographic data for Public Access datasets.

**Agriculture Plot Level Geographic Data**

All detailed geographic characteristics of plot level data (vertices, lines, polygons) should be removed from Public Access datasets, only retaining information at the administrative 2 (Admin 2) level. Prior to removing the geographic data, the plot polygon attributes should be generated by linking or spatially joining the polygon data (using geospatial tools) to the following geographic data sources:

* Soil characteristics data collected during the survey using LandPKS
* Agrometeorological data including rainfall, temperature, and soil moisture
* Other readily available geographic land features (land cover/use, hydrological, slope/topological, key biodiversity areas, depth to groundwater, etc.)

Advise your USAID/BFS Activity Manager if additional support is needed to link agrometeorological and other geospatial covariates to the ZOI Survey dataset.

The following attributes collected from interviews with plot managers and field data must be included:

* Plot manager description (sex, age, line number from household roster)
* Management practices and input use data (if more than one plot cultivated with the same crop, these data will be replicated across plots - farm-level data)
* Production estimates (if more than one plot cultivated with the same crop, these data will be replicated across plots - farm-level data)
* Total plot hectares calculated from plot geometry
* Distance from household to plot (\*calculate this prior to offsetting household latitude and longitude data described below)

Remove all geographic location details (except Admin 2) **after** the attributes for the plots have been created. Refer to Appendix E for an example of the Agriculture Module table.

**Household Level Geographic Data**

Public Access datasets will be shared with the public and will not contain the full geographic detail of the data collected. The following instructions are for offsetting the detailed household geographic data:

* + For each cluster, locate the approximate geographic center point based on locations of sampled households (excluding plots) and offset using the following guidelines:
    - Urban clusters contain a minimum of 0 and a maximum of 2 kilometers of error.
    - Rural clusters contain a minimum of 0 and a maximum of 5 kilometers of positional error with a further 1% of the rural clusters displaced a minimum of 0 and a maximum of 10 kilometers.

The offset should be restricted to remain within the country’s Admin 2 area boundary. If random offsets result in a geographic data point falling outside the relevant Admin 2 area’s boundary, its placement should be adjusted to remain within the Admin 2 boundary. For further guidance on how to offset geographic data points, refer to [The DHS Program - GPS Data Collection](https://dhsprogram.com/What-We-Do/GPS-Data-Collection.cfm) guidance.

Finally, remove all household location coordinates (latitude and longitude) **after** the distance from the household to the plot has been calculated and included in the agriculture plot data described above.

**NOTE:** While the restriction on named geographic areas and displacement and aggregation of granular geolocation data is important to respondent confidentiality for Public Access data, some data users may wish to analyze data at lower geographic levels. Partner organizations that have active awards with USAID and clauses of data confidentiality (non-disclosure agreement) included in the award may request access to the Restricted Access data sets that contain the granular geographic variables. These requests are handled on a case-by-case basis through USAID. USAID staff and interagency partners are also subject to data confidentiality clauses and non-disclosure agreements when requesting access to Restricted Access data. All requests for access to Restricted Access geocoded data files must be coordinated with USAID General Counsel/Regional Legal Officer or Authority, USAID Privacy, and the USAID project AO/CO and AOR/COR.

**5.2 Assess potential linkage to external datasets**

Surveys are often designed based on the use of existing sampling frames that form the basis of other surveys, creating a potential risk for linking data from one survey to those collected in another survey based on the same sampling frame. The household and individual identifier variables must be made anonymous, and all administrative location variables should also be anonymized, so that the data in the Zone of Influence Survey cannot be linked to other data sources. This ensures that there is no potential for respondent disclosure through linkage to other datasets.

**5.3 Manage direct and indirect identifier risks**

**Eliminate Direct Identifiers**

Per USAID guidance, the following respondent direct identifiers must be removed if collected:

* Names
* Remove or displace/offset and aggregate all geographic subdivisions/administrative units smaller than Admin 2 or area deemed too small to sufficiently manage risk of disclosure, including street address, city, county, precinct, ZIP code, and their equivalent geocodes and/or administrative names
* All elements of dates (except year) for dates that are directly related to an individual
* Ages > 89 years: the upper end of the age distribution should be top-coded appropriately; see the section on data value re-coding, below

**Manage High-Risk Indirect Identifiers**

Indirect identifiers are characteristics that alone do not identify a specific individual, but if used in combination with other information, may enable someone to identify an individual. Indirect identifiers may provide information that is critical to research studies in terms of demographics and other background information on study subjects and potentially prove valuable for further analysis. Removing them may reduce the usefulness of the dataset; however, their sensitivity should be assessed by determining the ability to use them in any combination to identify a unique individual.

Some of the most common indirect identifiers are background/demographic characteristics of people such as:

* Age
* Sex/Gender
* Marital status
* Race or ethnicity
* Employment or educational attainment

**NOTE:** The priority is to protect participants from disclosure; a user who has a need to replicate published data for follow-on analysis may apply for access to the Restricted Access dataset, and this may be granted with the necessary confidentiality pledge/non-disclosure agreement between the user organization and USAID. In these cases, the USAID AOR/COR will need to consult with USAID Privacy and General Counsel/Regional Counsel prior to sharing sensitive data. The syntax used to recode and/or suppress the original data should be included for analysts who have been granted restricted access to the data. Refer to the ***Registering and Submitting Data to the DDL*** guidance for more information.

**Methods to Manage High-Risk Indirect Indicators**

* **Data suppression** involves the removal of variables or cases from the data or setting specific values in specific cases to a missing value. It is the most effective in terms of disclosure avoidance but provides the public with data that are less usable. Due to this reduction in data usability, this approach is to be applied only to specific values where those values are outliers that cannot be combined with other responses. Such values will be set to missing with a numerical code that is not a valid value for the indicator, e.g. 9999 or 9998.
* **Data value re-coding** is to be applied to values that are discovered through frequency distributions or cross-tabulations to occur rarely enough to potentially enable disclosure. It involves re-coding existing values into new or existing groups. An example is a respondent’s age, if there are only a few people of a certain age. In this case, age could be re-coded as membership in an age group (such as 90 or more years old) instead of individual age. This prevents public users from duplicating report tables and indicators that have been posted by the project because the original values are no longer available, but it will render disclosure almost impossible.

**Step 1: Production of cross-tabulations for selected data items that may pose higher risk of identification in combination with other information**

Generate cross-tabulations to identify low-frequency, high disclosure risk variables.

The production of cross-tabulations should focus on two types of variables:

* Household characteristics readily discernible by casual observation; these could include household construction materials, size and composition of the household, and ownership of assets such as cars or livestock that set the household apart from others.
* Other respondent characteristics that might be known by a casual acquaintance, discovered through casual conversation or documented in other data sources. Examples of such characteristics include externally observable variables: anthropometric measurements, ethnicity, education, group membership or participation, and occupation.

**Step 2: Suppress or re-code high-risk data items**

Data values found to have been rarely used or outlier values should be cross-checked, where possible, with external data sources to confirm the rarity of the occurrence within each surveyed area. Viable external data sources that can serve this purpose include larger population-based surveys or national censuses.

If it is found that the rarely used or outlier values identified in Step 1 occur at a greater frequency in external data sources, then no modification to the ZOI Survey data is needed. If, however, those values do occur only rarely, for example 5 or fewer times at the district level, or if no viable external data source is available, then data modification will be required.

Values that are rarely found in the data will be recoded into groups to retain as much of the original information as possible while removing any possibility of a frequency cell with only 1 or 2 cases (for categorical variables), or only 1 or 2 cases at the upper or lower ends of the distribution (for continuous variables).

Outlier values are most often found in computed variables such as the Body Mass Index (BMI).  When discovered, these values will be re-coded to missing/flagged (NOTE: the DDL will not accept text values in numerical columns/cells. Use the standard numerical code for ZOI Survey data to represent missing values.)

**Step 3: Document the proposed approach to handling high-risk data items**

When cases are found that require either data suppression or recoding, prepare a table to document these findings (see Appendix C - Example Disclosure Analysis Plan) and provide the statistical programmer with sufficient information to implement suppression or re-coding. Each row of this table will represent a change to be made to the data. This table will include:

* **Data file**: the name or description of the data file in which the change is to be made
* **Variable name**: the name (not the label or description) of the variable requiring change
* **Variable description**: this is either the variable label or its general description
* **Variable type:** continuous or categorical
* **Value**: the existing value or values that require change
* **Changes**: a description of the modification to be made to the data; this may be a single value or a missing value, or the definition of the group into which the existing values will be re-coded. For example, “9 = 9+” could be used to mean, “recode all values 9 and higher as 9.”

**NOTE:** Remove all original values and their descriptions from the table created in Step 3 and include a copy of the table in the disclosure plan for the Public Access dataset.

**Step 4: Revision of codebook and other documentation**Certain details should be removed from supporting documentation to prevent the reconstruction of the original data files for the public access dataset. This will help prevent data users from uncovering the data edits that have been performed to protect respondent anonymity. The intentional omission of information must be acknowledged in supporting documentation such as a *Privacy Assurance Statement* and/or *Public Release README* file to accompany all datasets made available to the public. These files can be uploaded to the DDL registration form in the “Data Detail” tab under the “Other Reference Materials” section.

**Step 5: Label Files Appropriately**

Use the following naming convention example (or similar) for all Public Access dataset files to differentiate between Public, Restricted and Non-Public dataset files associated with a single Data Asset in USAID’s Development Data Library (DDL):

**EXAMPLE:** FTF \_ZOI\_2019\_Kenya\_ Household\_Public.csv

See Section VI below for more information on registering and submitting data to the DDL.

**NOTE:** The Data Codebook for the Public Access dataset must be revised to reflect the re-coding of categories and included as a Public Access file.

# VI. REGISTERING AND SUBMITTING DATA TO THE DEVELOPMENT DATA LIBRARY (DDL)

Implementing partners are required to meet their contractual obligations to submit their project and activity data to the [DDL](https://data.usaid.gov/) repository. The DDL enables partners to share data as widely as possible, while still protecting security and privacy. The DDL ensures that each submitter's data are thoroughly documented and screened for security and privacy risks, such as personally identifiable information when datafiles are to be made public. The DDL serves as a repository of valuable data from USAID projects and activities and makes them accessible to users with various levels of access. Much of the information documented following the *Protocol for preparing data for use* guidance can be copied and pasted to the DDL registration form fields and uploaded as supporting documentation.

**NOTE:**

Data Asset files fall into three access categories: Non-public access, Restricted access, and Public access.

**Non-public** access files are for USAID archival and use purposes. These access files fall under “principled exceptions to the presumption in favor of openness” established in OMB Bulletin 12-01, “Guidance on Collection of U.S. Foreign Assistance Data.” These exceptions include: When public disclosure threatens national security interests; When public disclosure is likely to jeopardize the personal safety of U.S. personnel or recipients of U.S. resources; When public disclosure would interfere with the Agency's ability to effectively discharge its ongoing responsibilities in foreign assistance activities; When there are legal constraints on the disclosure of business or proprietary information of non-governmental organizations, contractors, or private sector clients; When the laws or regulations of a recipient country apply to a bilateral agreement and restrict access to information; or When data reveal private information about individuals that must be kept confidential consistent with ethical guidelines and federal regulations (see ADS 508, Privacy Program).

**Restricted** access files may be shared with a user who has a need to replicate published data for follow-on analysis and has been granted access with the necessary confidentiality pledge/non-disclosure agreement between the user organization and USAID. In these cases, the USAID AOR/COR and AO/CO will need to consult with USAID Privacy and General Counsel/Regional Legal Officer prior to sharing sensitive data. Restricted access files require the documentation of the syntax used to prepare and create intermediate and final computed variables and indicators and methodology for re-coding and suppression of variables be included with the data submission so that internal statistical tests of differences over time and other analyses can be done.

**Public** Access files are data that can be shared openly and can be downloaded from the DDL without restrictions.

Dataset files for all levels of access can be associated with a single Data Asset. The benefit of this is that elements of the Data Asset metadata files created during registration on the DDL can be assigned to the different access level files without having to create a separate Data Asset for each level of access file (e.g. you will only have to enter the Survey Description once). To clearly identify which Data Asset files are for public, restricted or non-public access please follow the naming convention described in greater detail under the *Prepare data files and assemble all components* section below.

The informed consent and reference materials should be the same for all items in the submission and should only need to be submitted once regardless of the access level. However, the dataset, codebook, and a README file explaining changes should be different for each version of the data to ensure usability and reduce the potential risk of re-identification of respondents in Public Access data.  When submitting mixed-access Dataset files, ensure that you assign “**Public**” access under the Utility-Risk tab of the DDL registration form, and note in the *Proposed Access Level Comments* section of the form that you are submitting a combination (Restricted, Non-Public, and Public Access) of files.

**6.1 Request Account**

Partners must first request an account to register and submit data to the DDL if they have not done so previously. Follow these steps to create an account, register and submit data:

* Select **Register** or **Submit Data**from the [DDL homepage](about:blank) ([https://data.usaid.gov/](about:blank)). On the resulting screen, select **User Registration**.
* Next, select “**Partner Account”** to be taken to the form to request data submission rights on the platform. The form asks for information about you, the implementing partner organization submitting the data, and the partner organization’s relationship with USAID. Fill out this form completely—the Data Services team will use this information in an email to the project award AOR/COR. This email requests a response from the AOR/COR within five working days to verify that the individual requesting an account to upload data is a verified partner. If a request for editor rights is denied, the partner should contact their AOR/COR and/or the program office directly. The user can request further clarification regarding appropriate rights and permissions for submitting data to the DDL from their AOR/COR and/or the program office directly.

If your account is designated a verified partner, the Data Services team will grant you an editor role. You will receive an email notification that your role on the DDL has changed. Be sure to check your email account spam folder if you do not see the notification in your inbox. Once you have been given an editor role you can submit data to the platform immediately.

**6.2 Summary of Steps to Register and Submit Data**

* [**Prepare data files and assemble all Data Asset components**](https://data.usaid.gov/stories/s/Preparing-Data-for-Submission/2aex-zbcs):
* Convert to non-proprietary format: Data files must be converted to non-proprietary machine-readable formats such as csv. Other proprietary statistical software files should be included as well (e.g. SPSS, Stata).  Ensure data types are consistent throughout a column. The DDL will return error messages when data are not consistent throughout a column for example if the column uses a combination of text and numerical data. Ensure all text is in English[[1]](#footnote-1).
* Use consistent, standard naming conventions for all files. Include the level of access to be assigned to the data (Public, Restricted, Non-Public) at the end of the file name. Include the following common name parts in the Data Asset:

EXAMPLE: Feed the Future Cambodia Interim ZOI Survey (2015)

* + - USAID Program (Feed the Future)
    - Country Name (Cambodia)
    - Nature of the data collection (Interim ZOI Survey)
    - Year of data collection (2015)
  + For Datasets associated with the Data Asset, you can abbreviate common name parts to include the following:
    - USAID Program (FTF)
    - Country Name (Cambodia)
    - Nature of the data collection (Interim)
    - Year of data collection (2015)
    - Precise, short description of the dataset component (HouseholdMember\_Data, Children\_Data, Codebook, etc)
    - Access level (Public, Restricted)

EXAMPLES:

**Public Access:**FTF\_Cambodia \_Interim\_2015\_Children\_Public.csv

**Restricted Access:**FTF\_Cambodia\_Interim\_ 2015\_Children\_Restricted.csv

**Non-Public Access:**FTF\_Cambodia\_Interim\_ 2015\_Children\_NonPublic.csv

* Prepare supporting documentation and include the following files for the Public Access dataset:
* Upload de-identification/Disclosure Analysis Plan (see Protocol for preparing data use document) and Privacy Assurance Statement files describing steps taken to remove direct identifiers and mask high risk, indirect identifiers and ensure protection of privacy. These documents are uploaded in the “**Data Detail**” tab in the “**Other Reference Materials**” section.
* Copy of the Informed Consent (IC) form in English (no signatures or names included). This should be a stand-alone document and not embedded in the survey questionnaire.
* Public Release README file with information that facilitates proper use of the data by others. The file should include more detailed descriptions of each data file, a guide to linking the datasets, and instructions regarding characteristics of the data that should be considered during future analysis (such as the weighting of cases in a sample), and any use limitations or quality issues.
* Revised Codebook for the Public Access dataset (named with “Public” at the end of the file)
* Copy of the questionnaire in English, and all other languages to which the questionnaire was translated
* Other attachments such as links to the country report on the Development Experience Clearinghouse (DEC) or other intellectual reports
* Prepare supporting documentation (in addition to the Informed Consent and country report) and include the following files for the Restricted and Non-Public Access dataset files:
* Original or Revised Codebook (file name includes “Restricted” or “NonPublic” at the end)
* Syntax used to prepare and create intermediate and final computed variables and indicators
* [**Create and Register the Data Asset on the DDL**](https://data.usaid.gov/stories/s/Registering-Your-Data/t4bh-d9gr): A Data Assetis a group of Datasets with common characteristics (HH Member data file, WEAI data file, Household data file, etc.) that were all generated under the same survey. A Data Asset can share and automatically copy information (e.g. Survey Description) added to the metadata fields in the registration form to the Datasets it contains so you will not have to retype this information for each Dataset you associate with the Data Asset. Each Data Asset has a **primer page**, an overview of the metadata associated with the data asset and related datasets. Follow the naming convention recommendations under the *Preparing Data Files* section above and remember to include the access level at the end of the dataset names (Public, Restricted, NonPublic).
* **Register the Data Asset**: Complete all required fields with red asterisk and triangle icon (). You can copy and paste most of the information from the data documentation you created using the Protocol for preparing data for public use guidance to complete this tab. Include a brief description with the following information in the Data Overview tab (copy and paste from the Survey Description):
  + - A detailed description and extent of the geographic areas targeted by Feed the Future interventions, known as Feed the Future Zones of Influence (ZOI)
    - Names of contractor(s), sub-contractor(s)
    - Date range of when fieldwork took place
    - Total number of households interviewed for the survey
* [**Complete the Risk-Utility Assessment**](https://data.usaid.gov/stories/s/Risk-Utility-Assessment/r5k3-sb56): For each Data Asset all required fields (indicated by red asterisk and triangle () in the Risk-Utility Assessment section must be completed. Describe briefly in the “**Value of the Data”** section how the data can be used by other practitioners to do analyses, what problems and or research questions the data can help solve/answer, how the data will be used and if it fills any data gaps, temporal context, any data quality issues etc. In the “**De-Identifying**” section you can indicate that a supporting document exists and describes in detail how direct and indirect identifiers were removed, masked, anonymized, etc. The de-identifying methodology document should be uploaded in the “**Data Detail**” tab under the “**Other Reference Materials**” section. Answer Yes/No in the sections “Direct and Indirect Identifiers” and select the identifiers that apply. In the “**Proposed Access Level**” section select “**Public**” and provide comments under the Proposed Access Level Comments section briefly describing why the data asset can be made public – reference the de-identifying methodology document in the “Data Detail” tab.
* [**Add Datasets**:](https://data.usaid.gov/stories/s/Adding-Datasets/2jvq-bbpd) Once all the required metadata tabs () are completed you will be able to submit your Dataset files for ingest into the DDL. When you associate Dataset files with a “parent” Data Asset then the metadata already completed for the Data Asset can be assigned to the Dataset files. Ensure that the Dataset files are named using the recommended naming convention:
  + **Public Access:** FTF\_Cambodia \_Interim\_2015\_Children\_Public.csv
  + **Restricted Access:** FTF\_Cambodia\_Interim\_ 2015\_Children\_Restricted.csv
  + **Non-Public Access:** FTF\_Cambodia\_Interim\_ 2015\_Children\_NonPublic.csv
* [**Submit for Review**](https://data.usaid.gov/stories/s/Submitting-Data-for-Review/8hst-6dvg): When the metadata are complete and your data files and documentation have been uploaded, click on the Submit for Review button on the top right of the page. Check to make sure that your submission includes all the items noted above under preparing data files for submission. Select a visibility level for your submission: public or private. Select **Public** here, which will allow the USAID Data Services team to review your submission.

**NOTE:** Make sure you select “Public” for visibility level before clicking “Submit”. This is different from labeling the dataset files “Public” and will ensure that the USAID Data Service Team has access to the datasets to begin the review process.

* [**Review and Clearance Process**](https://data.usaid.gov/stories/s/Data-Clearance-Process/sqqm-yxhd): Once the data have been added and submitted to the DDL, the USAID Data Services team will determine if any information is missing or requires additional processing before the submission can be accepted. They will get in touch with you to request the missing or updated information. If all the needed information is present, Data Services will prepare an archival package of the accepted, original submission to facilitate preservation and ask the Operating Unit of Origin (Mission, Bureau etc.) to determine if your submission is an appropriate candidate for public release. Datasets labeled as Restricted or NonPublic will be archived and noted as such by Data Services. If the submission is designated appropriate for public use, Data Services will initiate the Clearance Process for all datasets labeled as Public. The clearance process can take several weeks.

# 

# APPENDIX A: CONTACTS AND RESOURCES

Bureau for Food Security Data Steward: Anna Brenes ([abrenes@usaid.gov](about:blank))

Feed the Future Population-Based Survey Methodology: Kiersten Johnson ([kiejohnson@usaid.gov](mailto:kiejohnson@usaid.gov))

USAID Data Services: [dataservices@usaid.gov](about:blank)

Development Data Library User Guide: ([https://data.usaid.gov/stories/s/ncfa-rh2w](about:blank))

Frequently Asked Questions About Open Data: ([https://data.usaid.gov/stories/s/7nq9-vptc](about:blank))

ADS 579 USAID Development Data: ([https://www.usaid.gov/sites/default/files/documents/1868/579.pdf](about:blank))

USAID Public Access Plan: ([https://www.usaid.gov/sites/default/files/documents/15396/USAID\_PublicAccessPlan.pdf](about:blank))

Geographic Data Collection and Submission Standards An Additional Help for ADS 579: (<https://www.usaid.gov/sites/default/files/documents/15396/579saa.pdf>)

The DHS Program’s Field Guide to GPS Data Collection: (<https://dhsprogram.com/publications/publication-dhsm9-dhs-questionnaires-and-manuals.cfm>)

# 

# APPENDIX B: EXAMPLE PRIVACY ASSURANCE STATEMENT

**PRIVACY ASSURANCE STATEMENT FOR THE**

**[YEAR] [COUNTRY] FEED THE FUTURE ZOI SURVEY**

**PUBLIC USE DATASET**

Under the terms of USAID's [Development Data Policy](https://www.usaid.gov/ads/policy/500/579) and [Public Access Plan](https://www.usaid.gov/open/public-access-plan), USAID is making data from Feed the Future population-based surveys available as open and machine-readable public use datasets after publication of the country reports.

An essential step in the preparation of datasets for public use is ensuring respondent privacy and confidentiality. Respondent privacy and confidentiality has been ensured and protected by removing and/or masking identifying information from the datasets, including *direct identifiers* (information such as names, addresses, Global Positioning System (GPS) coordinates, or any other personally identifying number or characteristic) and *indirect identifiers* (data that do not specifically identify a person or location, but that can be used to do so, one variable at a time or in combination, by uniquely describing a person or household). Parental permission was obtained prior to interviewing children under age 18. All local and national privacy protection laws (if applicable include a copy of the letter from host country government) were complied with.

Feed the Future datasets are freely available to the public without restriction. However, all personal identification information that may have been present in the original data has been removed. In addition, variables that could potentially be used indirectly (in combination with other variables or with other data sources) to identify any particular household or respondent have been re-coded to prevent that disclosure.

Because of these modifications it may not be possible for analysts to exactly reproduce all results in the published country reports and indicators.

# APPENDIX C: EXAMPLE DISCLOSURE ANALYSIS PLAN

**DISCLOSURE ANALYSIS PLAN**

**AND RECORD OF IMPLEMENTATION FOR THE**

**[YEAR] [COUNTRY] FEED THE FUTURE ZOI SURVEY**

**PUBLIC USE DATASET**

**This version of the disclosure analysis plan (DAP) is for USAID/BFS internal use only and is not intended for distribution outside of USAID.**

Under the terms of USAID's [Development Data Policy](https://www.usaid.gov/ads/policy/500/579) and [Public Access Plan](https://www.usaid.gov/open/public-access-plan), USAID is making data from Feed the Future’s Zone of Influence Surveys available as open and machine-readable public use datasets available on USAID’s [Development Data Library (DDL)](https://data.usaid.gov/).

An essential step in the preparation of datasets for public use is ensuring respondent privacy and confidentiality. Respondent privacy and confidentiality is ensured and protected by removing and/or masking identifying information from the datasets, including direct identifiers (information such as names, addresses, Global Positioning System (GPS) coordinates, or any other personally identifying number or characteristic) and indirect identifiers (data that do not specifically identify a person or location, but that can be used to do so, one variable at a time or in combination, by uniquely describing a person or household).

**Activity Name: [COUNTRY] 2025 ZOI Survey Public Use Datasets Preparation**

**Date: June 24, 2026**

**1. SURVEY DESCRIPTION**

Feed the Future Zone of Influence (ZOI) Surveys collect data through population-based household surveys to inform Feed the Future indicators and further analysis in the geographic areas targeted by Feed the Future interventions.

In [COUNTRY], the ZOI Survey fieldwork was conducted by [LOCAL SURVEY IMPLEMENTER] with technical assistance from [PRIME CONTRACTOR OR COOPERATING PARTNER]. The fieldwork took place from [DD/MM/YYYY] to [DD/MM/YYYY]. The ZOI in [COUNTRY] covers [XX NUMBER] districts in [four XX NUMBER] provinces ([NAME], [NAME], [NAME], and [NAME]). A total of [XX NUMBER] enumeration areas were selected from the sampling frame, from which [XXXX NUMBER] households were interviewed for the survey.

**2. STEPS TAKEN TO ENSURE COMPLIANCE WITH DATA USE RESTRICTIONS**

A review of relevant contracts and agreements was conducted. [PRIME CONTRACTOR OR COOPERATING PARTNER] has reviewed all agreements with USAID/BFS, IRB boards, as well as the survey questionnaire and informed consent statements to ensure that the public release datasets are in compliance with those agreements and statements.

Approval to conduct the ZOI Survey in [COUNTRY] was obtained on [DD/MM/YYYY] from the Republic of [COUNTRY], Ministry of Health, National Bioethics Committee for Health. The approval letter from National Bioethics Committee for Health gives [PRIME CONTRACTOR OR COOPERATING PARTNER] the full ethical clearance to collect and disseminate ZOI Survey data.

In addition, [PRIME CONTRACTOR OR COOPERATING PARTNER] required further clarification of the rights of children and a clear understanding of how their rights will be protected under local laws. As a result of this request, [LOCAL SURVEY IMPLEMENTER] obtained parental consent for the participation of their children under age 18 years. In households with no adults (no male or female age 18 or older), a household interview was conducted if there was a child age 15 who provided assent. Neither [LOCAL SURVEY IMPLEMENTER] nor local laws place any further restriction on the collection and dissemination of the survey data.

**3. ASSESSMENT OF LOCATION AND ADMINISTRATIVE VARIABLES**

MASKING OF GEOGRAPHIC INFORMATION

There are multiple variables collected in the survey that identify the geographic location of respondents. The 2025 [COUNTRY] ZOI Survey includes the location variables: cluster, stratum, A03 (Village), A04 (Administrative post), A05 (District), A06 (Province), and Urbrur (Urban Rural Indicator). These variables were necessary during data collection and data processing, but some of these variables also present a substantial risk to the anonymity of respondents. An assessment was performed to determine which location variables may undermine respondent confidentiality. Typically, only larger administrative areas, such as provinces or districts, will be identified by name in datasets, because an area must contain sufficient population and sample to minimize the risk that a household may be identified through indirect identifiers.

While the restriction on named geographic areas is important to respondent confidentiality, some data users may wish to analyze data at lower geographic levels. To facilitate this, variables at a lower geographic level will be anonymized by randomly renumbering the values for the variables. Anonymization will ensure that different respondents living in the same geographic area (such as a village) can be grouped together for analysis while protecting confidentiality by preventing the values for the geographic variable from being linked to other data sources (such as census data). For example, a village code that appears in the data does not correspond to the administrative codes used by the national government or the codes used by any other organization. Households within the same village are located in the same village in the dataset, but users cannot know which village it is. Without anonymization, it is possible that the values for the geographic variable may correspond to administrative codes (such as those in the census), which would allow a user to link the location variable with a variable in an external data source that would allow households to be identified.

All location variables have been anonymized except A06 (Province) and Urbrur.

ASSESSMENT OF POTENTIAL LINKAGE TO EXTERNAL DATASETS

Survey data are often collected using existing sampling frames that form the basis of other surveys. The household and individual identifier variables have been made anonymous by randomly renumbering the values for the variables, and because all administrative location measures (other than Province and Urbrur) have also been anonymized, the data in this survey cannot be linked to other data sources. This ensures that there is no potential for respondent disclosure by linkage to other datasets. The only potential linkage to external datasets is the matching of aggregated data results based on the location variable A06 (Province).

**4. MANAGEMENT OF DIRECT AND INDIRECT IDENTIFIERS**

ELIMINATION OF DIRECT IDENTIFIERS

All names of individuals and their addresses have been removed from the data. The day will be suppressed from all dates except the date of interview. In particular, dates of birth will include month and year, but not the day, which will greatly reduce the risk of identifying individuals. No telephone numbers or identification numbers were collected during this survey. Household GPS coordinates were offset using the following methodology:

For each cluster, the approximate geographic center point was located based on locations of sampled households (excluding plots) and offset as follows:

* + Urban clusters contain a minimum of 0 and a maximum of 2 kilometers of error.
  + Rural clusters contain a minimum of 0 and a maximum of 5 kilometers of positional error with a further 1% of the rural clusters displaced a minimum of 0 and a maximum of 10 kilometers.

The offset was restricted to remain within the country’s Admin 2 area boundary. If random offsets resulted in a geographic data point falling outside the relevant Admin 2 area’s boundary, its placement was adjusted to remain within the Admin 2 boundary.

MANAGEMENT OF INDIRECT IDENTIFIERS

**Step 1. Production of cross tabulations for selected data items.**

Cross tabulations were generated to identify low-frequency, high disclosure risk variables by the smallest geographic variable with geographic names in the datasets. For the 2025 [COUNTRY] ZOI Survey, this is the location variable A06 (Province).

The production of cross tabulations focused on two types of variables:

* Household characteristics readily discernible by casual observation; these would include household construction materials, and ownership of assets such as cars or livestock that set the household apart from others, and
* Other respondent characteristics that might be known by a casual acquaintance, discovered through casual conversation or documented in other data sources; examples of such characteristics include occupation and other economic activities.

For the [COUNTRY] 2025 ZOI Survey, the indirect disclosure risks discovered are summarized in Table 1. This table also indicates the proposed changes to be made to reduce disclosure risks.

**Step 2. Suppressing or recoding high-risk data items.**

Individual cases were identified as having outlier values for variables that were readily identified by observation of the physical household or by a casual acquaintance of the household. The values of these variables were then recoded to mask the outlier value while retaining the case in the data.

When cases were found that required either data suppression or recoding, they were inserted into a table to document these findings and provide the statistical programmer with sufficient information to implement suppression or recoding. For the 2025 [COUNTRY] ZOI Survey, the indirect disclosure risks identified in crosstabs were summarized and instructions were made for programmers to implement changes (see Table 1).

**Step 3. Revision of codebook and other documentation**

Certain details have been removed from the documentation to prevent the reconstruction of the original data files. This will help prevent data users from uncovering the data edits that have been performed to protect respondent anonymity. This intentional omission of information is acknowledged in the Privacy Assurance Statementand thePublic Release README, which should accompany all datasets made available to the public.

The Data Codebook and the Public Release README will contain, in addition to the information contained within this document, information to guide users of the Public Release Data. The codebook will be revised to reflect the re-coding of categories required by the DAP.

**Table 1: Summary of indirect risks and changes in data made to reduce the risks**

**(Province (A06): [NAME]=3, [NAME]=4, [NAME]=5, [NAME]=6)**

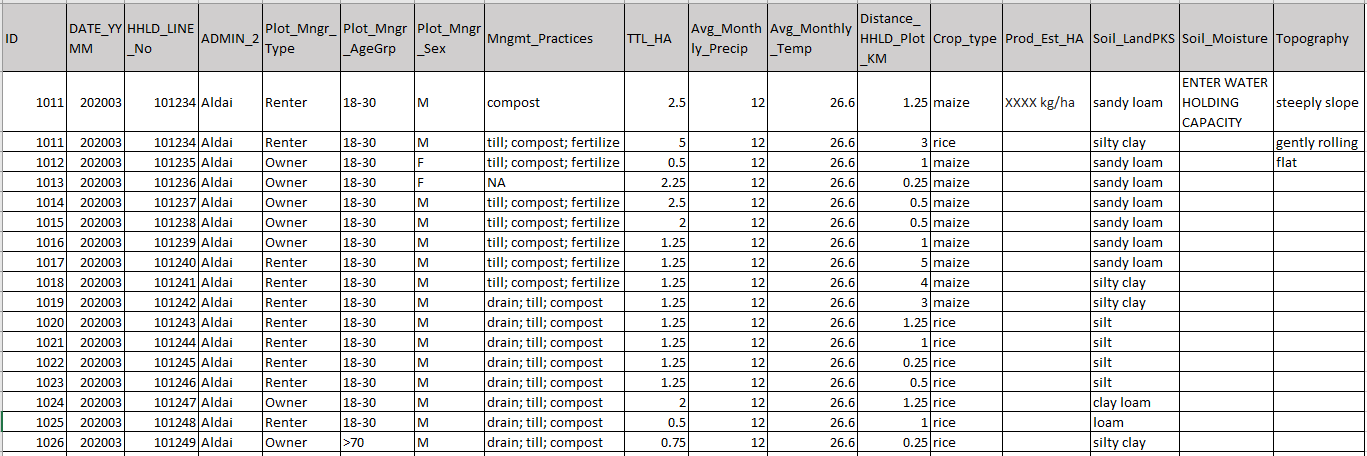
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **HOUSEHOLD Data file** | | | | |
| **Variable** | **Variable Description** | **Variable Type** | **Value** | **Changes** |
| d01 | Roof top material | Categorical | Natural roofing  No roof........................ 11  Other natural............. 19  Rudimentary roofing  Palm/bamboo.............. 22  Wood planks.............. 23  Cardboard................... 24  Plastic sheet................ 25  Other rudimentary... 29  Finished roofing  Ceramic tiles.............. 34  Cement....................... 35  Other finished........... 39 | create category “Other natural” with a value of 19  create category “Other rudimentary” with a value of 29  create category “Other finished” with a value of 39  recode d01=19 if d01=11  recode d01=29 if d01= 22, 23, 24, 25  recode d01=39 if d01=34, 35 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Variable Description** | **Variable Type** | **Value** | **Changes** |
| d03 | Exterior wall | Categorical | Natural walls  No walls.............. 11  Cane/Palm/  Trunks................. 12  Dirt..................... 13  Palm leaves........ 14  Other natural... 19  Rudimentary walls  Stone with mud...22  Uncovered adobe................... 23  Other rudimentary....... 29  Finished walls  Stone with lime/cement........ 32  Cement blocks................... 34  Wood planks/shingles.... 36  Fiber cement...... 37  Metal.................... 38  Other finished.... 39 | create category “Other natural” with a value of 19  create category “Other rudimentary” with a value of 29  create category “Other finished” with a value of 39  recode d03=19 if d03=11,12  recode d03=19 if d03=13 and a06= 5  recode d03=19 if d03=14 and a06= 4  recode d03=29 if d03=22 and a06=5, 6  recode d03=29 if d03=23 and a06=5  recode d03=39 if d03=32, 36, 37, 38  recode d03=39 if d03=34 and a06=6 |
| d04 | Number of rooms | Continuous | Number | recode d04=3 if d04>=3 and label this value “3+” |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **WEAI Data file** | | | | |
| **Variable** | **Variable Description** | **Variable Type** | **Value** | **Changes** |
| g301b\_a | Agricultural land | Continuous | Number | recode g301b\_a=4 if g301b\_a>=4 and label this value “4+” |
| g301b\_b | Large livestock | Continuous | Number | recode g301b\_b=1 if g301b\_b>=1 & g301b\_b<=4;  and label this value “1 to 4”  recode g301b\_b=5 if g301b\_b>=5; and label this value “5+” |
| g301b\_e | Fish pond / fishing equipment | Continuous | Number | recode g301b\_e=1 if g301b\_e>=1 and label this value “1+” |
| g301b\_g | Mechanized farm equipment | Continuous | Number | recode g301b\_g=1 if g301b\_g>=1 and label this value “1+” |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EXP\_E7 Data file** | | | | |
| **Variable** | **Variable Description** | **Variable Type** | **Value** | **Changes** |
| e701=368 | Satellite dish | Continuous | Number | recode e703=1 if e703>=1 & e701=368;  and label this value “1+” |
| e701=369 | Solar panel | Continuous | Number | recode e703=1 if e703>=1 & e701=369;  and label this value “1+” |

# APPENDIX D: EXAMPLE AGRICULTURE MODULE TABLE FOR PUBLIC ACCESS DATASETS



1. USAID does not have official guidance or policy on requiring data and its components to be in English. The partner should review the contract/grant which may contain a clause(s) requiring all deliverables be in English. USAID will not clear non-English data deliverables for public access. [↑](#footnote-ref-1)