Feed the Future

Survey Implementation

Document

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

Zone of Influence Surveys

Updated: April 2022

**Recommended Citation**

ICF. 2022. *Feed the Future Survey Implementation Document: Protocol for Preparing Non-Public, Restricted, and Public Access Datasets.* Surveys for Monitoring in Resilience and Food Security. Rockville. Maryland: ICF.

**Contact Information**

Feed the Future

1300 Pennsylvania Ave, NW

Washington, DC 20004

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

rfs.ald@usaid.gov

**TABLE OF CONTENTS**

[LIST OF ABBREVIATIONS iii](#_Toc102478970)

[1. INTRODUCTION 1](#_Toc102478971)

[2. DATASET TYPES 1](#_Toc102478972)

[2.1 Data files by dataset type 1](#_Toc102478973)

[2.2 Supporting documentation by dataset type 3](#_Toc102478974)

[3. ENSURING COMPLIANCE WITH DATA USE RESTRICTIONS 5](#_Toc102478975)

[4. INSTRUCTIONS FOR PREPARING NON-PUBLIC ACCESS DATASETS 5](#_Toc102478976)

[4.1 Assemble household geographic detail 5](#_Toc102478977)

[4.2 Assemble agriculture plot-level geographic data (as applicable) 6](#_Toc102478978)

[4.3 Address direct identifiers 6](#_Toc102478979)

[4.4 Create non-public access data files 6](#_Toc102478980)

[4.5 Create supporting documentation 7](#_Toc102478981)

[4.6 Label files appropriately 8](#_Toc102478983)

[5. INSTRUCTIONS FOR PREPARING RESTRICTED ACCESS DATASETS 8](#_Toc102478984)

[5.1 Assemble cluster-level geographic detail 8](#_Toc102478985)

[5.2 Assemble agriculture plot-level geographic data (as applicable) 8](#_Toc102478986)

[5.3 Address direct and indirect identifiers 9](#_Toc102478987)

[5.4 Create restricted access data files 9](#_Toc102478988)

[5.5 Create supporting documentation 10](#_Toc102478989)

[5.6 Label files appropriately 11](#_Toc102478990)

[6. INSTRUCTIONS FOR PREPARING PUBLIC ACCESS DATASETS 11](#_Toc102478991)

[6.1 Assemble and offset household geographic detail 11](#_Toc102478992)

[6.2 Remove agriculture plot-level geographic data (as applicable) 11](#_Toc102478993)

[6.3 Assess potential linkage to external datasets 12](#_Toc102478994)

[6.4 Manage direct and indirect identifier risks 12](#_Toc102478995)

[6.5 Create public access data files 12](#_Toc102478996)

[6.6 Create supporting documentation 13](#_Toc102478997)

[6.7 Label files appropriately 14](#_Toc102478998)

[7. SUBMITTING NON-PUBLIC ACCESS DATASETS 14](#_Toc102478999)

[8. REGISTERING AND SUBMITTING RESTRICTED AND PUBLIC ACCESS DATA TO THE DDL 15](#_Toc102479000)

[8.1 Prepare survey description 15](#_Toc102479001)

[8.2 Request account 15](#_Toc102479002)

[8.3 Summary of steps to register and submit data 15](#_Toc102479003)

[APPENDIX A: CONTACTS AND RESOURCES 19](#_Toc102479004)

[APPENDIX B: EXAMPLE PRIVACY ASSURANCE STATEMENT 20](#_Toc102479005)

[APPENDIX C: MANAGING HIGH-RISK INDIRECT IDENTIFIERS 21](#_Toc102479006)

[APPENDIX D: EXAMPLE DISCLOSURE ANALYSIS PLAN 25](#_Toc102479007)

# LIST OF ABBREVIATIONS

A-WEAI Abbreviated Women’s Empowerment in Agriculture Index

AOR Agreement Officer’s Representative

CAPI computer-assisted personal interviewing

COR Contracting Officer’s Representative

CSPro Census and Survey processing System

csv comma-separated value

DDL Data Development Library

GPS Global Positioning System

RFS Bureau for Resilience and Food Security

USAID United States Agency for International Development

ZOI Zone of Influence

# INTRODUCTION

Under the terms of the United States Agency for International Development’s (USAID) [Development Data Policy](https://www.usaid.gov/ads/policy/500/579) and [Public Access Plan](https://www.usaid.gov/open/public-access-plan#:~:text=USAID's%20Public%20Access%20Plan%20is,broad%20based%20development%20objectives%3B%20and), USAID is making data from Feed the Future’s Zone of Influence (ZOI) Surveys available as open and machine-readable public access datasets on USAID’s [Development Data Library (DDL)](https://data.usaid.gov/).

An essential step in the preparation of datasets for public access is ensuring respondent privacy and confidentiality. Respondent privacy and confidentiality is ensured and protected by removing or masking identifying information from the datasets, including *direct identifiers* (information such as names, 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).

This guidance document provides instructions to ZOI Survey implementers on (1) preparing non-public access datasets for submission to the Agreement Officer’s Representative/Contracting Officer’s Representative (AOR/COR) of the USAID ZOI Survey contract or agreement, (2) preparing restricted and public access datasets for submission to USAID’s DDL, and (3) creating and submitting documentation that provides information essential to using each dataset.

If ZOI Survey implementers using this protocol need additional information on how to prepare ZOI survey datasets, please refer to Appendix A for a list of additional contacts and resources.

# DATASET TYPES

USAID’s Bureau for Resilience and Food Security (RFS) requires partners who have implemented a Feed the Future ZOI Survey to submit three types of datasets, including a non-public access dataset to the AOR/COR of the USAID ZOI Survey contract or agreement as well as restricted and public access datasets to USAID’s DDL. Brief descriptions of these dataset types are as follows:

* A **non-public access** dataset retains some direct and many high-risk indirect identifiers.
* A **restricted access** dataset contains plot-level geocoded data and other sensitive indirect identifiers necessary for internal and external analysis.
* A **public access** dataset has been processed to remove all direct identifiers and suppress or remove high-risk indirect identifiers.

## 2.1 Data files by dataset type

For ZOI Surveys, the data files included in each type of dataset will differ. The data files needed for each dataset are summarized in Table 1.

Table 1. Summary of Data Files by Dataset Type

| **Dataset** | | |
| --- | --- | --- |
| **Non-public access**  **data files** | **Restricted access**  **data files** | **Public access**  **data files** |
| 1. Household | 1. Household | 1. Household |
| 2. Persons  including household members, primary adult decision-makers, women, children, and farmers | 2. Persons  including household members, primary adult decision-makers, women, children, and farmers | 2. Primary adult decision-makers |
| 3. Household members |
| 4. Women |
| 5. Children |
| 6. Farmers |
| 3. A-WEAI results | 3. A-WEAI results | 7. A-WEAI results |
| 4. Household location |  |  |
|  | 4. Cluster-level offset location | 8. Cluster-level offset location |
| 5. Agriculture plots\* | 5. Agriculture plots\* | 9. Agriculture plots\* |
| 6. Agriculture plot shape files\* | 6. Agriculture plot shape files\* |  |

\* As applicable

A-WEAI=Abbreviated Women’s Empowerment in Agriculture Index

Note: Data files may vary across surveys due to the content of country-customized survey questionnaires.

The public access dataset should have separate data files for (1) households,[[1]](#footnote-2) (2) primary adult decision-makers, (3) household members, (4) women of reproductive age, (5) children eligible for the nutrition module, (6) farmers, (7) agriculture plots, (8) the computed Abbreviated Women’s Empowerment in Agriculture Index (A-WEAI) results, and (9) offset cluster location, as relevant.[[2]](#footnote-3)

The non-public and restricted access datasets should consist of the following: (1) a household file with all household-level data; (2) a persons file with all person-level data, including data for primary adult decision-makers; (3) an agriculture plot file (as applicable); and (4) a file with the computed A-WEAI results.,[[3]](#footnote-4) Both the non-public and restricted access datasets should also include the agriculture plot shape files (if applicable), and the non-public access dataset should include a household location data file. Finally, as with the public access dataset, the restricted access dataset should include an offset cluster location file. Additional information on the dataset files to be included in each dataset can be found in Section 4.4 (Create non-public access data files), Section 5.4 (Create restricted access data files), and Section 6.5 (Create public access data files), as well as in the Feed the Future ZOI Survey Methods Toolkit[[4]](#footnote-5) in the Excel file titled “Example Variables for Feed the Future ZOI Survey Datasets.”

## 2.2 Supporting documentation by dataset type

The content contained in the supporting documentation for each dataset will also differ due to the information that is allowed in each type of dataset. In Sections 4-6 of this document, further detail on the preparation of the datasets, including what information is allowed under each dataset type, is provided.

Table 2 summarizes the required supporting documentation that should accompany the non-public, restricted, and public access dataset submissions. It also includes a brief overview of where and how to submit the data and accompanying files.

Table 2. Summary of Required Supporting Documentation by Dataset Type

| **Type of dataset** | **Where to submit** | **How to submit** | **Required supporting documentation** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Informed consent** | **QRE** | **READ-ME**  **file** | **Privacy assurance statement** | **Code-book** | **Analysis syntaxa** | **Recoding syntaxb** | **DAP** | **Final survey reportc** |
| **Public access** | DDL | (1) Create Data Asset in the DDL  (2) Complete DDL submission form  (3) Submit public access data and supporting documentation | X | X | X | X | X | X | - | X (redacted) | X |
| **Restricted access** | DDL | (1) Create zip file with restricted access data files and documentation  (2) Submit zip file under **“Other Reference Materials”** on the **Data Detail** tab of the DDL submission form for the public access dataset | - | X | X | - | X | X | X | - | - |
| **Non-public access** | AOR/ COR and USAID/ RFS | (1) Submit files to AOR/COR and USAID/RFS using a secure file transfer method | X | X | X | - | X | X | - | X | X |

QRE=questionnaire, CSPro=Census and Survey Processing System, DAP=disclosure analysis plan

a Syntax to create intermediate and final computed variables and calculate sample weighted indicators and other tabulated results

b Syntax to recode, suppress, remove data from the non-public access dataset

c Include the final 508-compliant survey report (or a link to the report if available on the Internet) only if the host country government has approved the report for public dissemination by the time of data submission to the DDL.

# ENSURING COMPLIANCE WITH DATA USE RESTRICTIONS

As a first step in preparing non-public, restricted, and public access datasets, it is important for ZOI Survey implementers to review their contract or agreement for relevant language on data use to ensure that the data and supporting documentation, including informed consent, Institutional Review Board approvals, etc., follow their contract or agreement clauses.

Approval to conduct the ZOI Survey in the country must have been obtained and documented prior to implementation of the survey. The approval letter from the host government 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 Institutional Review Board (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 18 years of age 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).

# 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” rule established in OMB Bulletin 12-01, “Guidance on Collection of U.S. Foreign Assistance Data.”[[5]](#footnote-6) Non-public access data are only available for internal use by the Federal Government, such as by a single program, single agency, or across multiple agencies.

The steps in the sections that follow should be followed to prepare the **non-public access** dataset and its supporting documentation.

## 4.1 Assemble household geographic detail

Non-public access datasets should include the following geographic detail in a standalone, but linkable, household location file:

* **Household latitude and longitude coordinates:** These data are essential for USAID to use in follow-on survey preparation and to inform evidenced-based analyses.

Offsetting location information is not required for the non-public access dataset.

## 4.2 Assemble agriculture plot-level geographic data (as applicable)

Some ZOI Surveys will collect agriculture plot-level data. For those surveys, the non-public access dataset should contain an agriculture plot shape file with the agriculture plot boundaries and attributes such as vertices, lines, and polygon geometry. For the polygon geometry, agriculture plot polygon attributes should be generated by linking or spatially joining the polygon data (using geospatial tools) to the following geographic data sources:

* Soil characteristic data (collected 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.)

**NOTE:** Contact the AOR/COR of the USAID ZOI Survey contract or agreement if support is needed to link agrometeorological and other readily available geospatial covariates to the ZOI Survey dataset.

In addition to agriculture plot polygon attributes in the agriculture plot shape file, the following attributes should be included in the agriculture plot data file in all three datasets:

* Total plot area in hectares calculated from plot geometry
* Distance from household to plot

Note that plot manager characteristics (e.g., sex, age, and household roster line number), management practices and input use data, and crop production estimates should be included in the farmers data file in the public access dataset and the persons data file in the restricted and non-public access datasets.

## 4.3 Address direct identifiers

ZOI Survey implementers should remove the following direct identifiers:

* Names

Note that ZOI Survey implementers do not need to remove indirect identifiers.

## 4.4 Create non-public access data files

The data files in the non-public access dataset may vary due to the content of country-specific survey questionnaires, but will generally include the following data files:

* **Household:** All household-level data, including data exported from the Census and Survey Processing System (CSPro) computer-assisted personal interviewing (CAPI) system and key intermediate and indicator variables. For ZOI baseline surveys, the household data file includes data related to household identification, Module 2 (Dwelling Characteristics), Module 3 (Food Security and Resilience), and Module 8 (Consumption Expenditure).
* **Persons:** All individual-level data, including data exported from the CSPro CAPI system and key intermediate and indicator variables. Note that Module 6 (Women’s Empowerment in Agriculture) data should be included in the persons data file even though they are in the household data file exported from CSPro. For ZOI baseline surveys, the persons data file includes data related to Module 1 (Household Roster and Demographics), Module 4 (Women’s Nutrition and Anthropometry), Module 5 (Children’s Nutrition and Anthropometry), Module 6 (Women’s Empowerment in Agriculture), and Module 7 (Agricultural Technologies), excluding plot-level data.
* **A-WEAI results:** A-WEAI overall results and by sex and age (e.g., all women, all men, women 18-29 years of age, men 18-29 years of age, women 30 years of age or older, men 30 years of age or older). The data file includes only key intermediate and indicator variables. The file is only included if the A-WEAI or its sub-indices (the Five Domains of Empowerment or the Gender Parity Index) were calculated.
* **Household location:** See Section 4.1 for details.
* **Agriculture plots:** See Section 4.2 for details. Included only if agriculture plot data were collected.
* **Agriculture plot shape files:** See Section 4.2 for details. Included only if agriculture plot data were collected.

With the exception of the agriculture plot shape files, each data file should be submitted in comma separated value (csv) and Stata formats.[[6]](#footnote-7) Variables relevant for each data file can be found in the Feed the Future ZOI Survey Methods Toolkits under the Excel file, “Example Variables for Feed the Future ZOI Survey Datasets,” which complements this document.[[7]](#footnote-8)

## 4.5 Create supporting documentation

Non-public access datasets should include the following supporting documentation:

* **Codebook:** Include the final survey-adapted codebook that includes key intermediate variables[[8]](#footnote-9) and all final indicator and disaggregate variables. All variables in the non-public access data files should be in the codebook submitted with the non-public access dataset.
* **Syntax:** Include the syntax used for creating the intermediate and final computed variables and calculating sample-weighted indicators and other tabulated results
* **Questionnaires:** Include the final questionnaires used in the field (all languages)
* **Copy of informed consent forms/statements used:** This should be a stand-alone document and not embedded in the survey questionnaire.
* **Other supporting documentation:**
  + README file providing details on each data file included in the non-public access dataset, including a description of the survey (see Section 8.1) and information on data use limitations[[9]](#footnote-10) and any other contextual information that facilitates use of the non-public access data by other data analysts
  + Disclosure analysis plan (see Appendix D for an example)

## 4.6 Label files appropriately

Use the following naming convention (or similar) for all non-public access dataset files:

* **EXAMPLE:** FTF\_Kenya\_ZOI\_2019\_Household\_Nonpublic.csv.

See Section 7 for more information on submitting non-public access datasets and files to the USAID AOR/COR of the ZOI Survey contract or agreement and USAID/RFS.

# 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 be included with the data submission. In addition, the syntax for recoding and suppressing variables should also be included with the data submission.

The steps below should be followed to prepare the **restricted access** dataset and its supporting documentation.

## 5.1 Assemble cluster-level geographic detail

Household latitude and longitude coordinates should not be included in the restricted access dataset. Rather, restricted access datasets should contain randomly offset cluster-level latitude and longitude data, which are generated from the household location data. This information should be contained in a standalone, but linkable, cluster-level location file. The data should be offset following the guidance in Section 6.1.

## 5.2 Assemble agriculture plot-level geographic data (as applicable)

As with non-public access datasets, restricted access datasets should contain the full agriculture plot‑level data (where applicable). This includes an agriculture plot shape file with agriculture plot boundaries and attributes such as vertices, lines, and polygon geometry. To assemble the agriculture plot‑level data, please follow the guidance provided in Section 4.2. However, be certain to remove all household location and plot location coordinates (latitude and longitude) from the data files **after** the distance from the household to the plot has been calculated and included with the agriculture plot data.

## 5.3 Address direct and indirect identifiers

The following direct and indirect identifiers, if included among the data collected, should be removed from the restricted access dataset:

* Names
* Geographic subdivisions/administrative units smaller than Admin 2 or other areas deemed too small to sufficiently manage risk of disclosure, including city, county, precinct, postal code, and their equivalent geocodes and/or administrative names
* Any unique identifying number or characteristic
* All elements of dates and ages (except year) that are directly related to an individual older than 6 years of age, including birth date (NOTE: include day, month, and year of birth for children 6 years of age or younger, and year only for all other respondents[[10]](#footnote-11))
* Ages >89 years: The upper end of the age distribution should be top-coded appropriately. Birth year should also be “bottom coded” so that ages >89 years are masked.

Instructions on data value recoding and other approaches to manage direct and indirect indicators are found in Appendices C and D.

## 5.4 Create restricted access data files

The data files in the restricted access dataset may vary across surveys due to the content of country-specific survey questionnaires but will generally include the following data files:

* **Household:** All household-level data, including data exported from the CSPro CAPI system and key intermediate and indicator variables. For ZOI baseline surveys, the household data file includes data related to household identification, Module 2 (Dwelling Characteristics), Module 3 (Food Security and Resilience), and Module 8 (Consumption Expenditure).
* **Persons:** All individual-level data, including data exported from the CSPro CAPI system and key intermediate and indicator variables. Note that Module 6 (Women’s Empowerment in Agriculture) data should be included in the persons data file even though they are in the household data file exported from CSPro. For ZOI baseline surveys, the persons data file includes data related to Module 1 (Household Roster and Demographics), Module 4 (Women’s Nutrition and Anthropometry), Module 5 (Children’s Nutrition and Anthropometry), Module 6 (Women’s Empowerment in Agriculture), and Module 7 (Agricultural Technologies), excluding plot-level data.
* **A-WEAI results:** A-WEAI overall results and by sex and age (e.g., all women, all men, women 18-29 years of age, men 18-29 years of age, women 30 years of age or older, men 30 years of age or older). The data file includes only key intermediate and indicator variables. The file is only included if the A-WEAI or its sub-indices (the Five Domains of Empowerment or the Gender Parity Index) were calculated.
* **Cluster-level offset location:** See Section 5.1 for details.
* **Agriculture plots:** See Section 5.2 for details. Included only if agriculture plot data were collected.
* **Agriculture plot shape files:** See Section 5.2 for details. Included only if agriculture plot data were collected.

With the exception of the agriculture plot shape files, each data file should be submitted in csv and Stata formats. The data files should be prepared such that indirect and direct identifiers are addressed according to the instructions in Sections 5.1-5.3. Variables relevant for each data file can be found in the Feed the Future ZOI Survey Methods Toolkits under the Excel file, “Example Variables for Feed the Future ZOI Survey Datasets,” which complements this document.[[11]](#footnote-12)

## 5.5 Create supporting documentation

Restricted access datasets should include the following supporting documentation:[[12]](#footnote-13)

* **Codebook:** Include the restricted access codebook, which amends the non-public access codebook to reflect all adjustments made to the variables or value sets for the restricted access dataset
* **Syntax:** Include both the analysis syntax used for creating intermediate and final computed variables and calculating sample-weighted indicators and other tabulated results and the recoding syntax used to recode, suppress, or remove data from the non-public access dataset.
* **Questionnaires:** Include the final questionnaires used in the field (all languages)
* **Other supporting documentation:**
  + README file providing details on each data file included in the restricted access dataset, including a description of the survey (see Section 8.1) and information on data use limitations[[13]](#footnote-14) and any other contextual information that facilitates use of the restricted access data by other data analysts

## 5.6 Label files appropriately

Use the following naming convention (or similar) for all restricted access data files:

* **EXAMPLE:** FTF\_Kenya\_ZOI\_2019\_Household\_Restricted.csv

See Section 8 for more information on registering and submitting data to the DDL.

# INSTRUCTIONS FOR PREPARING PUBLIC ACCESS DATASETS

**Public access datasets** are or could be made publicly available without restrictions. All direct identifiers collected during the survey must be removed or sufficiently anonymized to protect the privacy of respondents. 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 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 details on how to prepare the public access dataset while mitigating and managing risk to ensure that the data can be made public.

## 6.1 Assemble and offset household geographic detail

Public access datasets will be shared with the public and will not contain the full geographic detail of the data collected. As such, individual household GPS coordinates will not be included in public access datasets. The following instructions should be used to offset the detailed household geographic data for the public access dataset:

* For each cluster, using the household location data file in the non-public access dataset, locate the approximate geographic center point based on locations of sampled households (excluding plots). Offset the data using the following guidelines:
  1. Urban clusters contain a minimum of 0 and a maximum of 2 kilometers of error.
  2. Rural clusters contain a minimum of 0 and a maximum of 5 kilometers of positional error with a further 1 percent 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.
* Include the randomly offset cluster-level location data in a standalone, but linkable, cluster location file for submission with the public access dataset.

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.

## 6.2 Remove agriculture plot-level geographic data (as applicable)

Do not submit agriculture plot shape files with the public access dataset. Keep geographic data at the Admin 2 level and higher. Be certain to remove all household location and plot location coordinates (latitude and longitude) from the agriculture plot data file included in the public access dataset after the distance from the household to each plot has been calculated and included in the agriculture plot data. An example of a public access agriculture plot data file is included in the Feed the Future ZOI Survey Methods Toolkits under the Excel file titled “Example Variables for Feed the Future ZOI Survey Datasets.”

## 6.3 Assess potential linkage to external datasets

Surveys are often designed using 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. For the public access dataset, the household and individual identifier variables as well as the Admin 2 variables should be anonymized. This ensures that there is no potential for respondent disclosure through linkage to other datasets.

## 6.4 Manage direct and indirect identifier risks

For the public access dataset**,** the following respondent direct and indirect identifiers **must be removed:**

* Names
* Geographic subdivisions/administrative units smaller than Admin 2 or other areas deemed too small to sufficiently manage risk of disclosure, including city, county, precinct, postal code, and their equivalent geocodes and administrative names
* Any unique identifying number or characteristic
* All elements of dates (except year) that are directly related to an individual[[14]](#footnote-15)
* Ages >89 years: The upper end of the age distribution should be top-coded appropriately. Birth year should also be “bottom coded” so that ages >89 years are masked.

Instructions on data value recoding and other approaches to manage direct and indirect indicators are found in Appendices C and D.

## 6.5 Create public access data files

The data files in the public access dataset may vary across surveys due to the content of the country-customized survey questionnaires, but the public access dataset will generally include the following data files:

* **Household:** All household-level data, including data exported from the CSPro CAPI system and key intermediate and indicator variables. For ZOI baseline surveys, the household data file includes data related to household identification, Module 2 (Dwelling Characteristics), Module 3 (Food Security and Resilience), and Module 8 (Consumption Expenditure).
* **Household member:** All individual-level data related to Module 1 (Household Roster and Demographics), including data exported from the CSPro CAPI system and key intermediate and indicator variables.
* **Women:** All data for women of reproductive age, including data exported from the CSPro CAPI system and key intermediate and indicator variables. For ZOI baseline surveys, the women’s data file includes data related to Module 4 (Women’s Nutrition and Anthropometry).
* **Children:** All data for children, including data exported from the CSPro CAPI system and key intermediate and indicator variables. For ZOI baseline surveys, the children’s data file includes data related to Module 5 (Children’s Nutrition and Anthropometry).
* **Primary adult decision-makers:** All data for primary adult decision-makers, including data exported from the CSPro CAPI system and key intermediate and indicator variables. For ZOI baseline surveys, the primary adult decision-maker data file includes data related to Module 6 (Women’s Empowerment in Agriculture).
* **A-WEAI results:** A-WEAI results overall and by sex and age (e.g., all women, all men, women 18-29 years of age, men 18-29 years of age, women 30 years of age or older, men 30 years of age or older). The data file includes only key intermediate and indicator variables. This file is only included if the A-WEAI or its sub-indices (the Five Domains of Empowerment or the Gender Parity Index) were calculated.
* **Cluster-level offset location:** See Section 6.1 for details.
* **Farmers:** All data for farmers responsible for making management decisions about value chain livestock or crops selected for inclusion in the survey, including data exported from the CSPro CAPI system and key intermediate and indicator variables. For ZOI baseline surveys, the farmers data file includes data collected in Module 7 (Agricultural Technologies).
* **Agriculture plots:** See Section 6.2 for details. Included only if agriculture plot data were collected.

Each data file for the public access dataset should be submitted in csv format. The data files should be prepared such that indirect and direct identifiers are addressed according to the instructions in Sections 6.1-6.4. Variables relevant for each data file can be found in the Feed the Future ZOI Survey Methods Toolkits under the Excel file, “Example Variables for Feed the Future ZOI Survey Datasets,” which complements this document.[[15]](#footnote-16)

## 6.6 Create supporting documentation

Information on how direct and indirect identifiers were addressed in the public access dataset should be removed from supporting documentation to prevent the reconstruction of the original data files. This will help prevent data users from uncovering data edits that have been performed to protect respondent anonymity. The intentional omission of information must be acknowledged in supporting documentation, such as theprivacy assurance statement or public accessREADME file to accompany all datasets made available to the **public.**

Public access datasets should include the following supporting documentation:

* **Codebook:** Include the public access codebook, which amends the non-public access codebook to reflect any adjustments made to the variables or value sets for the public access dataset
* **Syntax:** Include the syntax used for creating intermediate and final computed variables and calculating sample-weighted indicators
* **Questionnaire:** Include the final questionnaires used in the field (all languages)
* **Copy of informed consent forms/statements used:** Include as a stand-alone document and not embedded in the survey questionnaire
* **Other supporting documentation:**
  + Public access README file providing details on each data file included in the public access dataset, including a description of the survey (see Section 8.1) and information on data use limitations[[16]](#footnote-17) and any other contextual information that facilitates use of the public access data by other data analysts
  + Privacy assurance statement (see Appendix B for an example)
  + Disclosure analysis plan. The disclosure analysis plan included with the public access dataset is a redacted version of the disclosure analysis plan submitted to USAID. As noted in Appendix D, the redacted version of the disclosure analysis plan removes all original values and their descriptions from the table created in Step 3.
  + Other relevant files, such as the publicly available survey reports, created using the survey data. If the reports are publicly available, the links to the reports on the Development Experience Clearinghouse are sufficient to include in the supporting documentation.

## 6.7 Label files appropriately

Use the following naming convention example (or similar) for all public access dataset files to differentiate between public and restricted access dataset files associated with a single data asset in USAID’s DDL:

* **EXAMPLE:** FTF\_Kenya\_ZOI\_2019\_Household\_public.csv

See Section 8 for more information on registering and submitting data to the DDL.

# SUBMITTING NON-PUBLIC ACCESS DATASETS

ZOI Survey implementing partners are required to use a secure file transfer method (e.g., encrypted zip files, Kiteworks) to submit the non-public access dataset and supporting documentation, including the questionnaires, codebook, syntax for creating intermediate and final computed variables and calculating sample-weighted indicator estimates, the disclosure analysis plan, README file, and final survey report (if available) to their USAID AOR/COR and to USAID/RFS. Prior to submitting, discuss recommended file transfer options with your USAID AOR/COR or USAID/RFS point of contact.

# REGISTERING AND SUBMITTING RESTRICTED AND PUBLIC ACCESS DATA TO THE 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 data files 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. The descriptions in the sections that follow provide step-by-step instructions for registering and submitting the public and restricted access datasets to the DDL.

## 8.1 Prepare survey description

In preparation for submitting restricted and public access datasets to the DDL, ZOI Survey implementers should develop a detailed description of the survey for the datasets. This information can be used to copy and paste to the DDL registration form data asset metadata tabs described in further detail as follows. The description should include the following information:

* A detailed description and extent of the geographic areas targeted by Feed the Future interventions, known as Feed the Future ZOIs
* Names of contractors and local survey implementation partners
* Date range of when fieldwork took place
* Total number of households interviewed for the survey

**EXAMPLE:** In [COUNTRY], the [NAME OF SURVEY] fieldwork was conducted by [SURVEY SUBCONTRACTOR], with technical assistance from [CONTRACTOR]. The fieldwork took place from [DATE - DD/MM/YYYY] to [DATE - DD/MM/YYYY]. The ZOI in [COUNTRY] at the time of the survey covered [XX NUMBER] districts in [XX NUMBER] regions. A total of [XX NUMBER] enumeration areas were selected from the sampling frame, from which [XXXX NUMBER] households were interviewed for the survey.

## 8.2 Request account

Partners will need to create a partner account to register and submit data to the DDL if they have not done so previously. Please follow the steps for [Creating an Account](https://data.usaid.gov/stories/s/Creating-an-Account/y88r-z9wv), located in the [DDL User Guide](https://data.usaid.gov/stories/s/ncfa-rh2w).

## 8.3 Summary of steps to register and submit data

1. **[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.
* Ensure the data type is consistent for each variable (i.e., ensure that throughout a column in each csv file, the data type in all cells is consistent). The DDL will return error messages if data are not consistent throughout a **column** (e.g., if the column uses a combination of text and numerical data).
* Ensure that all text is in English.[[17]](#footnote-18)
* Use consistent, standard naming conventions for all files. Include the level of access to be assigned to the data (public or restricted) at the end of the file name.

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

* + USAID program: Feed the Future (acceptable abbreviation: FTF)
  + Country name: Cambodia
  + Geography of the data collection: ZOI
  + Date: 2015
  + Precise, short description of the dataset component: HouseholdMember, Children, Codebook, etc.
  + Access level: Restricted, public

**Public access:** FTF\_Cambodia\_ZOI\_2015\_Children\_public.csv

**Restricted access:** FTF\_Cambodia\_ZOI\_2015\_Children\_restricted.csv

1. [**Create and register the data asset on the DDL**](https://data.usaid.gov/stories/s/Registering-Your-Data/t4bh-d9gr)**:** Each that 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 Step 1 (Prepare data files) and remember to include the access level at the end of the dataset names (public, restricted).

* **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 supporting documentation you already created to complete this tab. Copy and paste the survey description that you created in Section 8.1 in the **Data Overview** tab.

1. [**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 perform analyses, what problems and or research questions the data can help solve or answer, how the data will be used and whether they fill any data gaps, describe the temporal context, describe 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 **Direct and Indirect Identifiers** sections 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. In the **Proposed Access Level Comments** section, it is also helpful to note that the restricted access dataset and supporting documentation has been included as a zip file in the **Data Detail** tab under **Other Reference Materials** section.

**NOTE:** The restricted access dataset will only be made available to users upon request. Requests are made using an [Access Request Form](https://data.usaid.gov/access-request) on the DDL and are handled on a case-by-case basis through USAID. USAID staff and interagency partners may request access to restricted data on the DDL through a [ServiceNow form](https://usaiditsm.servicenowservices.com/sphome?id=sc_cat_item&sys_id=ecce569ddb6d7b00e2c0fd0e0f9619ac&sysparm_category=1c1b4f18db6d7700e2c0fd0e0f961955), selecting the category “Restricted Data Access Request.” USAID staff and interagency partners are also subject to data confidentiality clauses and non-disclosure agreements when requesting access to restricted access data.

1. **Request embargo for Feed the Future ZOI Survey data assets:** All endline Feed the Future ZOI Survey data assets (including public and restricted access datasets, final reports, and any other data products that present study findings) must be embargoed until after USAID publishes the Progress Snapshot in any given year.

Please see the following table for embargo dates:

|  |  |
| --- | --- |
| **Report and data set posted:** | **Embargo until:** |
| October 1, 2021–September 29, 2022 | September 30, 2022 |
| October 1, 2021–September 29, 2023 | September 30, 2023 |
| October 1, 2022–September 29, 2024 | September 30, 2024 |

Embargo dates must be entered in the DDL registration under the **Embargo Section** tab. Provide the following rationale: “We are requesting an embargo due to the pending publication of USAID’s official Progress Snapshot.”

1. [**Add datasets:**](https://data.usaid.gov/stories/s/Adding-Datasets/2jvq-bbpd) After all the required metadata tabs () are completed, you will be able to submit your dataset files for ingestion 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

1. [**Submit for review**](https://data.usaid.gov/stories/s/Submitting-Data-for-Review/8hst-6dvg)**:** When the metadata are complete and 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 previously under preparing data files for submission. Select a visibility level for your submission: public or private.

**NOTE:** Be sure to select “**public**” for visibility level before clicking on **Submit.** This is different from labeling the dataset files “public” and will ensure that the USAID Data Services team has access to the datasets to begin the review process.

7. [**Review and clearance process**](https://data.usaid.gov/stories/s/Data-Clearance-Process/sqqm-yxhd)**:** After the data have been submitted to the DDL, the USAID Data Services team will determine whether any information is missing or requires additional processing before the submission can be accepted. They will contact you to request any missing or updated information. If all the needed information is present, the USAID Data Services team will prepare an archival package of the accepted, original submission to facilitate preservation and ask the Operating Unit of Origin (e.g., Mission or Bureau) to determine whether your submission is an appropriate candidate for public release. Datasets labeled as restricted will be archived and noted as such by the USAID Data Services team. For submissions that are designated appropriate for public use, the USAID Data Services team will initiate the clearance process. The clearance process can take several months to complete.

# APPENDIX A: CONTACTS AND RESOURCES

Feed the Future Population-Based Survey Methodology: Anne Swindale ([aswindale@usaid.gov](mailto:aswindale@usaid.gov))

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

Development Data Library User Guide: (<https://data.usaid.gov/stories/s/ncfa-rh2w>)

Frequently Asked Questions About Open Data: (<https://data.usaid.gov/stories/s/7nq9-vptc>)

ADS 579 USAID Development Data: (<https://www.usaid.gov/sites/default/files/documents/579.pdf>)

USAID Public Access Plan: (<https://www.usaid.gov/open/public-access-plan#:~:text=USAID's%20Public%20Access%20Plan%20is,broad%20based%20development%20objectives%3B%20and>)

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

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

Example variables for Feed the Future P2-ZOI Survey baseline DDL datasets:

* Feed the Future ZOI Survey Methods Toolkit - Baseline (2019) Toolkit Item 4.2

(<https://www.agrilinks.org/post/feed-future-zoi-survey-methods>)

* Feed the Future ZOI Survey Methods Toolkit – Midline (2021) Toolkit Item 4.1

(<https://www.agrilinks.org/post/feed-future-zoi-survey-methods-toolkit-midline-2021>)

# 

# APPENDIX B: EXAMPLE PRIVACY ASSURANCE STATEMENT

**PRIVACY ASSURANCE STATEMENT FOR THE**

**[YEAR] [COUNTRY] FEED THE FUTURE ZONE OF INFLUENCE SURVEY**

**PUBLIC ACCESS DATASET**

Under the terms of the United States Agency for International Development’s [Development Data Policy](https://www.usaid.gov/ads/policy/500/579) and [Public Access Plan](https://www.usaid.gov/open/public-access-plan), the Agency 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 or masking identifying information from the datasets, including *direct identifiers* (information such as names, Global Positioning System 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 18 years of age. 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 recoded to prevent that disclosure.

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

# APPENDIX C: MANAGING HIGH-RISK INDIRECT IDENTIFIERS

Indirect identifiers are characteristics that alone do not identify a specific individual, but, if used in combination with other information, they 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 or demographic characteristics of people, such as the following:

* Age
* Sex or 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 the United States Agency for International Development (USAID). In these cases, the USAID Agreement Officer’s Representative/Contracting Officer’s Representative will need to consult with USAID Privacy and General Counsel/Regional Counsel prior to sharing sensitive data. The syntax used to recode or suppress the original data should be included for analysts who have been granted restricted access to the data. The Risk Utility Assessment portion of the Data Development Library (DDL) submission form asks that you “Please describe any efforts you have already taken to de-identify potentially sensitive data within this dataset. If this is captured in a document, please upload this under the "Data Detail" tab under "Other Reference Materials," providing guidance on the submission of this information. Refer to Section 8, ***Registering and Submitting Restricted and Public Access Data to the DDL,*** guidance for more information.

For Zone of Influence (ZOI) Surveys, some indirect identifiers, such as location and age, will always need to be recoded in or excluded from restricted and public access datasets (see Table C1), whereas other variables may need to be recoded or excluded on a survey-to-survey basis (see the “Method to Manage High Risk Indirect Identifiers” section in this appendix).

Table C1 summarizes adjustments that should be made to age and location variables in all ZOI Survey restricted and public access datasets. Note that depending on survey-specific adjustments made to the questionnaire and indicators calculated, variables may vary among ZOI Surveys. Therefore, the adjustments specified in Table C1 apply to any similar survey-specific variables. No adjustments need to be made to age or location variables in non-public access datasets.

Table C1. Place, Age, and Location Variable Adjustments

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Survey module** | **Type of variable** | **Example ZOI Survey variables** | **Change for restricted access** | **Change for public access** |
| HHID | Admin 1 | a03d | (keep same) | (keep same) |
| Admin 2 | a03c | (keep same) | Randomize and anonymized |
| Admin 3 | a03b | Drop | Drop |
| Roster | Age | v104 | Top-code if person >89 years old | Top-code if person >89 years old |
| Children’s nutrition | Child’s day of birth | v502d, v506d | (keep same) | Drop |
| Child’s month of birth | v502m, v506m | (keep same) | Drop |
| Child’s year of birth | v502y, v506y | (keep same) | (keep same) |
| Child’s age in days | cage\_days | (keep same) | Drop |
| Child’s age in months | v508, cage\_months, cage\_months\_int | (keep same) | Drop |
| Child’s age in years | v104, v507 | (keep same) | (keep same) |
| Women’s nutrition | Woman’s month of birth | v402m | Drop | Drop |
| Woman’s year of birth | v402y | (keep same) | (keep same) |
| AWEAI | Primary adult decision-maker’s age | v6102, m6102 | Top-code if person >89 years old | Top-code if person >89 years old |
| Primary adult decision-maker’s month of birth | v6101m, m6101m | Drop | Drop |
| Primary adult decision-maker’s year of birth | v6101y, m6101y | Bottom-code if person >89 years old | Bottom-code if person >89 years old |
| HHID | Latitude | latitude | Randomly offset at cluster level | Randomly offset at cluster level |
| Longitude | longitude | Randomly offset at cluster level | Randomly offset at cluster level |
| LandPKS | Latitude | latitude (of soil pit) | Drop | Drop |
| Longitude | longitude (of soil pit) | Drop | Drop |

**Method to manage high-risk indirect identifiers**

* **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 in which 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 recoding** 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 recoding 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 recoded 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: Produce 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 that are 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 recode 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 in 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 five 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 one or two cases (for categorical variables), or only one or two cases at the upper or lower ends of the distribution (for continuous variables).

Outlier values are most often found in computed variables, such as Body Mass Index. When discovered, these values will be recoded 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 D, Example Disclosure Analysis Plan) and provide the statistical programmer with sufficient information to implement suppression or recoding. Each row of this table will represent a change to be made to the data. This table will include the following:

* **Data file:** Name or description of the data file in which the change is to be made
* **Variable name:** Name (not the label or description) of the variable requiring change
* **Variable description:** Either the variable label or its general description
* **Variable type:** Continuous or categorical
* **Value:** Existing value or values that require change
* **Changes:** 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 recoded. 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 analysis plan for the public access dataset.

# APPENDIX D: EXAMPLE DISCLOSURE ANALYSIS PLAN

**DISCLOSURE ANALYSIS PLAN**

**AND RECORD OF IMPLEMENTATION FOR THE**

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

**PUBLIC ACCESS DATASET**

**This version of the disclosure analysis plan is for United States Agency for International Development (USAID)/Bureau for Resilience and Food Security 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 (ZOIs) available as open and machine-readable public use datasets available on USAID’s [Development Data Library](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 masking identifying information from the datasets, including direct identifiers (information such as names, Global Positioning System 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: [YEAR] [COUNTRY] ZOI Survey Public Access Dataset Preparation**

**Date: [Submission Month Day, Year]**

**1. SURVEY DESCRIPTION**

Feed the Future 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 and relevant institutional review 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 [ETHICS COMMITTEE IN SURVEY COUNTRY] 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 18 years of age. In households with no adults (no male or female 18 years of age or older), a household interview was conducted if there was a child 15 years of age 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 [YEAR] [COUNTRY] ZOI Survey data include the location variables: cluster, stratum, a03a (Village), a03b (Administrative post), a03c (District), a03d (Province), and ahtype (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.

Although 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 a03d (Province) and ahtype (urban/rural designation).

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 a03d and ahtype) 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 a03d (Province).

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

ELIMINATION OF DIRECT IDENTIFIERS

All names of individuals 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 only the year, but not the month or day, which will greatly reduce the risk of identifying individuals. No telephone numbers or identification numbers were collected during this survey. Household Global Positioning System 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 percent 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 [YEAR] [COUNTRY] ZOI Survey, this is the location variable a03d (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.
* 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 [YEAR] [COUNTRY] ZOI Survey, the indirect disclosure risks discovered are summarized in Table D1. This table also indicates the proposed changes to be made to reduce disclosure risks.

**Step 2. Suppression or recoding of 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 [YEAR] [COUNTRY] ZOI Survey, the indirect disclosure risks identified in crosstabs were summarized, and instructions were made for programmers to implement changes (see Table D1).

**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 statement and the public access README file, which should accompany all datasets made available to the public.

The codebook and the public access README file will contain, in addition to the information contained within this document, information to guide users of the public access data. The codebook will be revised to reflect the recoding of categories required by the disclosure analysis plan.

Appendix D Table 1. Example of Summary of Indirect Risks and Changes in Data Made to Reduce the Risks

**(Province (a03d): [NAME]=3, [NAME]=4, [NAME]=5, [NAME]=6)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Household data file** | | | | |
| **Variable** | **Variable description** | **Variable type** | **Value** | **Changes** |
| v201 | 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 v201=19 if d01=11  Recode v201=29 if d01= 22, 23, 24, 25  Recode v201=39 if d01=34, 35 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Variable description** | **Variable type** | | **Value** | **Changes** |
| v203 | 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 v203=19 if v203=11,12  Recode v203=19 if v203=13 and a06= 5  Recode v203=19 if v203=14 and a06= 4  Recode v203=29 if v203=22 and a03d=5, 6  Recode v203=29 if v203=23 and a03d=5  Recode v203=39 if v203=32, 36, 37, 38  Recode v203=39 if v203=34 and a03d=6 |
| v204 | Number of rooms used for sleeping | Continuous | | Number | Recode v204=3 if v204>=3 and label this value “3+” |
| v8703\_36 | Satellite dish | | Continuous | Number | Recode v8703\_36=1 if v8703\_36>=1  and label this value “1+” |
| v8703\_37 | Solar panel | | Continuous | Number | recode v8703\_37=1 if v8703\_36>=1 & e701=369  and label this value “1+” |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Primary adult decision-maker data file** | | | | |
| **Variable** | **Variable description** | **Variable type** | **Value** | **Changes** |
| v6301\_01 | Agricultural land | Continuous | Number | Recode v6301\_01=4 if v6301\_01>=4 and label this value “4+” |
| v6301\_02 | Large livestock | Continuous | Number | Recode 26301\_02=1 if v6301\_02>=1 & g301b\_b<=4  and label this value “1 to 4”  Recode v6301\_02=5 if v6301\_02>=5 and label this value “5+” |
| v6301\_05 | Fishpond / fishing equipment | Continuous | Number | Recode v6301\_05=1 if v6301\_05>=1 and label this value “1+” |
| v6301\_08 | Mechanized farm equipment | Continuous | Number | Recode v6301\_08=1 if v6301\_08>=1 and label this value “1+” |

1. Consumption expenditure data (as relevant) should be included in the household data file. [↑](#footnote-ref-2)
2. Data files may vary across surveys. For example, if a community questionnaire is administered, the data should be saved in a separate community data file, or if plot data are not collected, the agriculture plot file is not needed [↑](#footnote-ref-3)
3. As with the public access dataset, the data files in the restricted and non-public access datasets may vary from survey to survey. For example, if a community questionnaire is administered, the data should be saved in a separate community data file [↑](#footnote-ref-4)
4. [Feed the Future ZOI Survey Methods Toolkit - Baseline (2019)](https://www.agrilinks.org/post/feed-future-zoi-survey-methods) and [Feed the Future ZOI Survey Methods Toolkit - Midline (2021)](https://www.agrilinks.org/post/feed-future-zoi-survey-methods-toolkit-midline-2021) [↑](#footnote-ref-5)
5. For the Feed the Future ZOI 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. [↑](#footnote-ref-6)
6. Note that csv and Stata data files can be read into other statistical analysis software packages such as R, SAS, and SPSS. [↑](#footnote-ref-7)
7. [Feed the Future ZOI Survey Methods Toolkit - Baseline (2019)](https://www.agrilinks.org/post/feed-future-zoi-survey-methods) Toolkit Item 4.2 and [Feed the Future ZOI Survey Methods Toolkit - Midline (2021)](https://www.agrilinks.org/post/feed-future-zoi-survey-methods-toolkit-midline-2021) Toolkit Item 4.1. [↑](#footnote-ref-8)
8. Key variables include those used to create indicator variables and those that are tabulated in the results tables. [↑](#footnote-ref-9)
9. For non-public access datasets, data use limitations include those related to sample design (e.g., the survey is meant to be representative at the ZOI level, so disaggregated estimates may be imprecise due to small sample sizes) and survey specific issues (e.g., a skip or filter was incorrectly set up in the data collection applications, so an indicator cannot be calculated according the to Feed the Future Indicator Handbook). [↑](#footnote-ref-10)
10. In the core Feed the Future ZOI Surveys, this includes respondents to the women’s nutrition module and the women’s empowerment in agriculture module. [↑](#footnote-ref-11)
11. [Feed the Future ZOI Survey Methods Toolkit - Baseline (2019)](https://www.agrilinks.org/post/feed-future-zoi-survey-methods) Toolkit Item 4.2 and [Feed the Future ZOI Survey Methods Toolkit - Midline (2021)](https://www.agrilinks.org/post/feed-future-zoi-survey-methods-toolkit-midline-2021) Toolkit Item 4.1. [↑](#footnote-ref-12)
12. Note that supporting documentation required for the restricted access dataset does not include the final questionnaires or copy of informed consent forms/statements used, or syntax used to create the intermediate and final computed variables and indicators. ZOI Survey implementers will submit these documents only with the public and non-public access datasets. [↑](#footnote-ref-13)
13. For restricted access datasets, data use limitations include those related to recoding and suppressing data, sample design (e.g., the survey is meant to be representative at the ZOI level, so disaggregated estimates may be imprecise due to small sample sizes), and survey specific issues (e.g., a skip or filter was incorrectly set up in the data collection applications, so an indicator cannot be calculated according the to Feed the Future Indicator Handbook). [↑](#footnote-ref-14)
14. Binary variables that indicate whether a child is 0-5 months of age and 6-23 months of age, and therefore eligible for inclusion in the children’s dietary intake indicator calculations, can be retained. [↑](#footnote-ref-15)
15. [Feed the Future ZOI Survey Methods Toolkit - Baseline (2019)](https://www.agrilinks.org/post/feed-future-zoi-survey-methods) Toolkit Item 4.2 and [Feed the Future ZOI Survey Methods Toolkit - Midline (2021)](https://www.agrilinks.org/post/feed-future-zoi-survey-methods-toolkit-midline-2021) Toolkit Item 4.1. [↑](#footnote-ref-16)
16. For the public access dataset, data use limitations include those related to recoding and suppressing data, sample design (e.g., the survey is meant to be representative at the ZOI level, so disaggregated estimates may be imprecise due to small sample sizes), and survey specific issues (e.g., a skip or filter was incorrectly set up in the data collection applications, so an indicator cannot be calculated according the to Feed the Future Indicator Handbook). [↑](#footnote-ref-17)
17. USAID does not have official guidance or policy on requiring data and their components to be in English. The partner should review the contract/grant, which may contain a clause requiring all deliverables be in English. USAID will not clear non-English data deliverables for public access. [↑](#footnote-ref-18)