USAID
FROM THE AMERICAN PEOPLE

DDL ROADMAP
FOR EDUCATION PROGRAMS

Get Started

WHY SUBMIT TO THE DDL?

USAID considers data a public good that should be shared with a global audience, while safeguarding privacy, security, and confidentiality. As expressed in its Operational Policy, USAID encourages the broad use of data by partners, academic and scientific communities, and the public at large. Making data accessible, discoverable, and usable fuels innovation, improves program design and implementation, and ultimately, empowers the students, schools, and communities we serve.

USAID established the Development Data Library (DDL) to serve as the official data repository for the agency. Through the DDL, partners can disseminate datasets and data assets, while secondary users can discover and download data for their own benefit. The data you submit to the DDL undergo a clearance review to ensure that USAID makes as much of the data publicly available as is practical, while securing confidentiality and safety of Agency staff and program beneficiaries.

This roadmap is designed to help partners in education fulfill their contractual obligation to submit activity data to the DDL. Beyond covering the requirements for submission, we offer guidance to make sure your data are of a sufficient quality to be discoverable and useful. By submitting high-quality data to the DDL, you can help fuel innovation, drive good practice, and brighten the future for the next generation of learners.

BRIGHTEN THE FUTURE FOR THE NEXT GENERATION

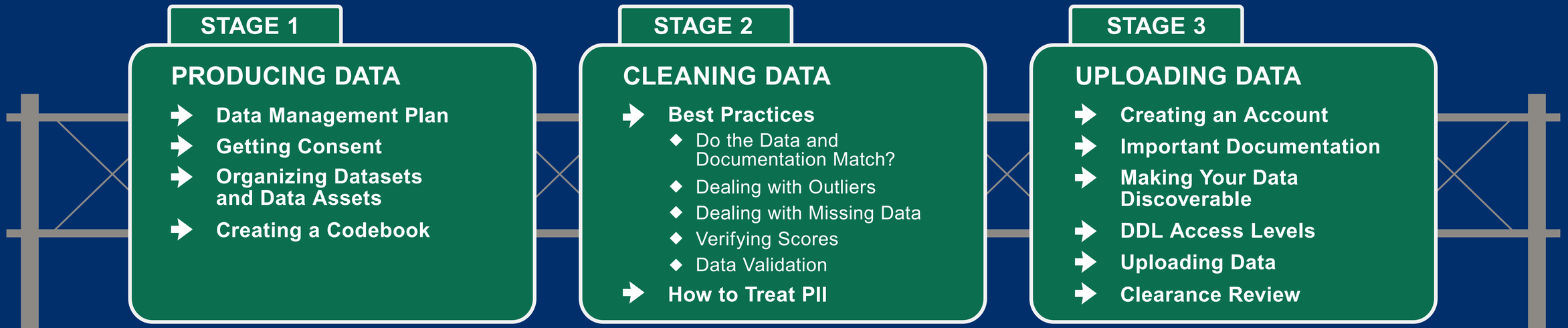
USAID invests in education because it is a fundamental driver of development. Throughout this roadmap, you will see icons indicating how your DDL submission will fuel innovation, drive good practice, and empower students, schools, and communities.

Guidance explaining how you can successfully submit to the DDL so others can discover and use your data.

Guidance explaining how you can strengthen the quality of your data so others can draw valid conclusions from them.

Guidance explaining how you can respect the rights and privacy of program beneficiaries.

DDL ROADMAP
FOR EDUCATION
PROGRAMS



HOW TO USE THE ROADMAP

This roadmap is your guide for submitting data to the DDL. The PDF is designed to be interactive--please feel free to navigate your journey using the icons and links on each screen.

In the following sections, you will find guidance uniquely tailored to partners in education, alongside links to articles and videos in the DDL User Guide, which is aimed at a broader audience of all DDL users and submitters. The parts of the roadmap that link to the User Guide are marked with the heading, "Explore the DDL User Guide Online."

ROAD SIGN LEGEND

STOP | This is a critical part of DDL submission. Your submission will not move into the clearance review without this element.

GREEN LIGHT | This is a best practice that will strengthen your submission.

CAUTION | This is advice for dealing with important issues that may come up when preparing your submission.

EXPRESS | This will speed up your understanding of how to navigate the DDL.

LINK LEGEND

Link to the DDL User Guide

Link to DDL User Guide Videos

Link to a glossary term

Link to an external resource

Link to a roadmap section

STAGE 1: PRODUCING DATA

DATA MANAGEMENT PLAN GO

To help you collect data in a way that complies with the requirements of your award, USAID is developing guidance for partners on data management planning. A data management plan (DMP) is a document that describes the data you intend to collect and how you will store, analyze, and share those data. Crafting a DMP with your Agreement/Contracting Officer's Representative (AOR/COR), as part of the Monitoring, Evaluation, and Learning (MEL) Plan, is one of the first steps you should take when planning an activity. USAID is working to establish requirements for DMPs for partners. In the meantime, the guidance we share below and throughout this roadmap will help you manage data responsibly, in line with your contractual obligation and USAID's expectations.

You can find more guidance on data management in the recently released Considerations for Using Data Responsibly, developed by USAID's Global Development Lab. The USAID Data Services team (dataservices@usaid.gov) is also available to help you with data management and design of a DMP.

BEST PRACTICES FOR DATA MANAGEMENT

- Create and maintain an inventory of the datasets and documentation that are required deliverables under your award provisions and guidelines. (See Organizing Datasets and Data Assets and Important Documentation for more information.)
- Make sure you are able to collect, store, and manage data responsibly, and identify any digital technology or software you will need to understand and use your data.
- Create rich documentation (including metadata and codebooks) that describe the data. (See Making Your Data Discoverable and Creating a Codebook for more information.)
- Make sure you are able to document and manage any privacy and security risks associated with the data. (See How to Treat PII and DDL Access Levels for more information.)
- Make sure any data-related legal agreements and informed consent documentation specify how the data can be accessed and shared. (See Getting Consent for more information.)
- Document and describe your procedures, including timelines, for submitting data and related documentation to a USAID-managed or approved data repository, such as the DDL.
- Document your plan, including timelines, for working with data repository experts, such as the Data Services team, to share and publish as much of your data as is practical, while ensuring confidentiality and safety of Agency staff and program beneficiaries.

These best practices are suggestions and are not mandatory. You should be able to adopt these practices without increasing your level of effort or costs; if not, please reach out to your AOR/COR and Agreement/Contracting Officer (AO/CO).

GETTING CONSENT

Informed consent is a critical part of data collection. A copy of the language used to obtain consent is also a critical part of your DDL submission (see Important Documentation for other required documents). Your submission cannot advance to the clearance review without the consent form attached. In the DDL User Guide resources linked on this page, you can find instructions on how to include this documentation in your submission.

USAID asks partners to get informed consent from all participants in a data collection activity, like an assessment or survey.[1] **This is to ensure that participants made an informed, voluntary decision to take part in the activity and understand that the data you collect may be shared with others.** When submitting to the DDL, you must include the language used to obtain consent, translated into English if necessary.[2] Data submitted without informed consent, or documentation from an Institutional Review Board (IRB) or Ethics Review Committee (ERC) explaining why consent was not required, might not be considered appropriate for publication on the DDL.[3] This, in turn, would hinder USAID's ability to share your data.

This section discusses considerations for informed consent that uniquely affect USAID's education partners. We discuss the language to include in the consent form, how to get informed assent from children, and other issues to consider. By following this guidance, you can help protect the communities you serve while making sure your data can advance to the clearance review.

EXPLORE THE DDL USER GUIDE ONLINE

How to attach a copy of the consent form to your DDL submission?

The data you submit to the DDL must include informed consent (or documentation from an IRB or ERC explaining why consent was not needed). The links below explain how to attach a copy of the consent form in the Data Detail tab when registering a dataset or data asset.

Registering Your Data

Data Detail

1 The exceptions are if the activity does not involve human participants or if all the data collected are already publicly available.

2 If the consent language is included in the instrument, and required to continue collecting data, the instrument can be used as the informed consent documentation for your submission.

3 Instead, the data's access level would be set to Non-Public. While secondary users would be able to view the metadata and primer page, they would not be able to access the actual data. See DDL Access Levels for more information.

(GETTING CONSENT, CONT.)

WHAT TO INCLUDE IN THE CONSENT FORM

The Code of Federal Regulations (CFR) has detailed instructions for the language to include in the informed consent form. In the education sector, some activities, like student testing and teacher training, may be exempt from the CFR, meaning you may not be required by law to obtain consent for those activities.[4] However, USAID asks partners to get informed consent for all activities, even those exempt from the CFR. The Agency asks that the consent form include, at the very least:

- The activity's purpose
- What you are asking participants to do in the activity
- How long you expect the person's participation in the activity to last
- Any risks or discomfort the person might experience while participating
- Any benefits the person might expect to receive from participating
- Whether you will collect personally identifiable information (PII), such as names and addresses, and if so, how you will keep the data private
- Whether you intend to share the data you collect with others outside your team, and whether you intend to share the data in a data repository, such as the DDL[5]
- Whom participants can contact with questions about the activity
- A statement that participation in the activity is voluntary, those who do not participate will not be punished, and a person can stop participating at any time

Ultimately, it is up to the IRB or ERC to decide whether the consent language is adequate. We encourage you to document the decisions the IRB or ERC makes regarding informed consent, and include that documentation in your submission along with a copy of the consent form itself.

GETTING CONSENT/ASSENT FROM CHILDREN

Especially in the education sector, activities often involve children younger than the age of majority and take place in schools where parents may not be immediately available. Because children are not legally able to give consent, you will need to decide whether to get informed consent from parents or whether it is enough to simply ask the children if they agree to participate (this is known as *informed assent*). An IRB or ERC will need to make this decision based on the laws in the jurisdiction where you operate and whether the activity is exempt from the CFR. Please note, however, that if your activity is not exempt from the CFR, getting informed assent from children is not enough; you will still need to get informed consent from parents.

OTHER CONSIDERATIONS

Other issues to consider when getting consent include:

- **Verbal consent**. In some cases, it may be better to get consent verbally instead of using a written form. A person may be illiterate, for example, or signing forms may not be a cultural norm. Moreover, if the signed form is the only record linking a person to your activity, and the person's participation puts them at some risk, then verbal consent would be a safer option.

4 You can find more information about exempt activities in Section 225.104 of the Code.

5 While we encourage you to share as much data as is practical, you should not share information that threatens the personal safety of participants. Nor should you share information that is so sensitive that people would not have participated in your activity had they known you intended to share the information. Indeed, we encourage you to reconsider collecting data in the first place if doing so threatens the safety of others.

(GETTING CONSENT, CONT.)

- **Appropriate words**. Please pay attention to the words used in the consent form, because some words may not make sense to everyone. Children and parents in rural areas, for example, may not understand words like "data" or "risk." Please use words and explanations participants can clearly understand.
- **Translations**. Please make sure the consent form is accurately translated when working in non-English speaking communities. You will want to make sure you are conveying the information Having someone who is fluent in the local language back-translate the form can help make sure it is accurate. However, the consent documentation you submit to the DDL must be in English to be eligible for clearance at the Public access level (see DDL Access Levels for more information).

STAGE 1:
PRODUCING DATA

STAGE 2:
CLEANING DATA

STAGE 3:
UPLOADING DATA

GLOSSARY

ORGANIZING DATASETS AND DATA ASSETS

In the DDL, datasets are grouped into data assets. A data asset acts as a "shell" that contains datasets related to one another. Before uploading a dataset to the DDL, you will first need to create the data asset that will contain it. Even if you are only uploading a single dataset, you will need to create a data asset first.

There are several ways to organize datasets and data assets. For example, if you conducted an evaluation of a reading activity, a data asset might contain datasets for EGRA results and teacher interviews. Or, if you conducted a baseline and an endline, you can group the datasets into the same data asset. **The way you group datasets into data assets can help secondary users spot trends and draw connections.** For example, it can make it easier to compare student-level and teacher-level data, or an endline and a baseline.

In the DDL User Guide resources linked on the right, you can find more guidance for organizing datasets and data assets. You will also find tips for creating useful titles and descriptions for your datasets and data assets. This will help others discover and understand the contents of your submission (see Making Your Data Discoverable for other tips to make your data easier to find).

GO

EXPLORE THE DDL USER GUIDE ONLINE

How to organize datasets and data assets?

Learn best practices for grouping datasets into data assets.

How to create useful titles and descriptions?

Learn best practices for creating titles and descriptions that are brief, precise, and informative.

Preparing Data for Submission

Navigating the DDL

CREATING A CODEBOOK

A codebook is an essential tool to help secondary users decipher your data. **Without a codebook, others might not understand what a variable represents or could misinterpret a variable's values.** For example, they might assume the value "99" means the numeric value 99, when instead, it may represent a missing value. A codebook is required for each dataset submitted to the DDL (see Important Documentation for other required documents).

The codebook should list the name and description of each variable in your dataset. As you create the codebook, please try to keep the variable names short and meaningful. Using an entire survey question to name a variable, for example, or using meaningless names like a1, a2, and so on, can make it difficult for others to work with your data. Please also avoid duplicating variables. For example, if you have one variable that reports participants' age and another that reports age ranges (e.g., 10-14, 15-19), you can remove one from the dataset (and codebook).

In the DDL User Guide resources linked on the right, you can find more information about what to include in the codebook and how to attach this document to your submission.

STOP

EXPLORE THE DDL USER GUIDE ONLINE

What to include in the codebook?

How to create a codebook that describes your data accurately and completely.

Preparing Data for Submission

How to include the codebook in your submission?

You can attach the codebook in the Data Detail tab when registering a dataset.

Registering Your Data

Data Detail

STAGE 2: CLEANING DATA

BEST PRACTICES

Low-quality data can be difficult, and sometimes impossible, to use. Submitting low-quality data to the DDL would undermine USAID's goal of fueling innovation and driving good practice. **The data cleaning tips we share in this section will help make sure your data are of a sufficient quality to be useful.** These best practices build on the Early Grade Reading Assessment (EGRA) Toolkit and other guidance for learning assessments, but are also largely applicable to surveys, observations, and other types of data.

Throughout the data cleaning process, it is important to document your efforts and explain what you are doing. This will make it easier for secondary users to understand your data. When registering a dataset or data asset, please attach this documentation as a README.txt under "Other Reference Materials" in the Data Detail tab. Also, please keep in mind that if you use Microsoft Excel formulas for data cleaning, these formulas will not be transferred when you upload your data to the DDL platform (see Uploading Data for more information).

DO THE DATA AND DOCUMENTATION MATCH?

When you submit data to the DDL, we encourage you to include data collection plans, evaluation reports, and other documentation that can aid secondary users. However, please make sure that the data and accompanying documentation correspond to each other. For example, if your data collection plan explains that you sampled 20 students per school in 10 schools, your dataset should have about 200 student records. If there is a mismatch or correspondence error, you should investigate and explain the reason (for example, you may not have been able to reach some schools). Here are some common things to look for at this step:

- **Sampling.** Do the data match the sampling plan? For example, if you intended your sample to be half boys and half girls, does the dataset have a roughly equal number of records for boys and girls?
- **Dates and times.** Were the data collected during the official data collection period? Data collected at odd times could indicate a trial-run or mistake and should be removed.
- **Data collection instrument.** Do the data match the instrument? For example, if you used a survey to collect data from head teachers, does your submission actually contain responses from head teachers? See the clipboard on the next page for other things to check when looking at the instrument.

DEALING WITH OUTLIERS

Outliers, or extreme values, can reflect errors. Below are some common sources of outliers and ways to address them.

- **Data entry errors.** The data could have been entered incorrectly. You may be able to correct this by contacting the data producer or cross-referencing the information with other data sources.

(BEST PRACTICES, CONT.)

- **Intentional misreporting.** The data may have been changed intentionally by the data collector or manager. Your team should have protocols in place to protect the integrity of the data and make sure no unauthorized manipulation occurs.
- **Data calculation errors.** Scores and other variables that are aggregated or calculated manually could contain errors. You can trace them by reviewing how the variable was calculated.
- **Legitimate cases.** It is possible for an outlier to be legitimate and not the result of an error. In a learning assessment given to low-performing students, for example, a few high scores might be considered outliers. In such cases, you should use a consistent approach to identify and deal with these values. For example, you may decide to remove any value that is three standard deviations above and below the mean. Or, you may decide to include any outlier so long as it is legitimate. In either case, you should explain your approach in the README.txt file, so that secondary users can understand.

DEALING WITH MISSING DATA

Missing data can confuse secondary users and delay the clearance review if USAID officials cannot determine what the values mean. For example, a common practice is to enter "99" for a missing record. Without an explanation of what "99" means, others could mistakenly interpret the observation as having the numeric value 99. This is one reason it is important to create a codebook that clearly explains what the values in a dataset mean (see Creating a Codebook for more information).

VERIFYING SCORES

If your submission includes data from a learning assessment, please check whether the scores were calculated correctly. Here are some common mistakes when calculating different types of scores:

- **Raw score.** This is the number of correct responses. If students did not attempt any items, they are sometimes given a score of zero; instead, the records should be coded as missing.

Things to look for when comparing the data and data collection instrument:

- **Number of questions.** Does the dataset have the same number of questions as the instrument?
- **Skip pattern.** To make data collection easier, instruments often use skip patterns, where respondents skip to another question or page based on their answer choice. A common mistake is to code observations for skipped items as missing data.
- **Administration time.** You should look for any period of time administering an instrument that seems suspicious (e.g., a learning assessment that took longer than others).

(BEST PRACTICES, CONT.)

- **Percent correct score.** This is the number of correct responses divided by the total number of items. Please check whether percent correct scores are inadvertently missing for some tasks. As with raw scores, students might mistakenly be given a score of zero if they did not attempt any items.
- **Time-adjusted scores.** For timed tasks, this score is the number of correct responses divided by the time taken to complete the task, then multiplied by 60 (assuming the maximum time is 60 seconds). For example, if a student pronounced 20 sounds correctly in 30 seconds, the time-adjusted score would be 20 ÷ 30 × 60 = 40. You should only adjust scores for time if a student completed all items in the task in less than 60 seconds.

DATA VALIDATION

Data validation is a continuous process of checking for errors (and correcting them). Below are some common errors to check when cleaning your data:

- **Correspondence errors.** Does your dataset contain missing or extra data? As discussed above, please check whether the data correspond to the data collection plan, and clearly label any missing records.
- **Invalid values.** Do any values seem unreasonable, such as a negative value for an assessment score or a very early or late survey start time?
- **Logic checks.** As discussed above, instruments sometimes use skip patterns. For example, a survey might include the question, "Does your school have a library?" followed by the question, "How many books are in the library?" Please check whether the responses make sense. If a respondent answered "no" to the first question, they should have skipped the second question. If the skip pattern was not followed correctly, however, a reply of "no" might be followed by a response like "100 books," which would not make sense.
- **Missing codes.** Are all codes explained in the codebook? For example, if responses to the variable *gender* are coded "1" for male and "2" for female, please check whether these codes are included and explained in the codebook.

HOW TO TREAT PII

When it comes to collecting and sharing data, there is always a tough trade-off between utility and privacy. On the one hand, we want to include enough details to make data useful. **On the other, we want to avoid sharing PII that could be used to identify a specific person.** In this section, we offer guidance for treating PII in your DDL submission to help you find the right balance between utility and privacy.

For some types of PII, the choice is clear. You should never include direct identifiers in your submission. Direct identifiers are data that can identify a specific person without any other details. Names, real-life ID numbers or school education management information system (EMIS) codes, and geographic information system (GIS) coordinates are good examples. Please remove or mask these before submitting to the DDL.

For indirect identifiers, the situation is more complex. Indirect identifiers are data that may not, alone, identify a person, but could identify someone when combined with other information. Age, gender, and smaller geographic divisions, such as a district or city, are good examples. Indirect identifiers can be useful for analysis. For example, smaller geographic divisions are sometimes used for survey weighting. Students' age and gender are also important variables when analyzing education data. However, indirect identifiers also affect privacy risk--the risk of identifying a specific person. For example, age could be used to identify an individual if some participants are much older or younger than others.

To reduce privacy risk, there are strategies to de-identify data. For example, you can replace individual ages with broader age ranges (e.g., 10-14, 15-19, etc.). See the clipboard on the right for common de-identification

Common de-identification strategies:

- **Use broader categories.** You can replace individual ages with age ranges, for example.
- **Top-code or bottom-code.** You can replace the highest or lowest values with averages. For example, if for age you have 6, 7, 8, 8, and 12, and 12 is an outlier, you can replace the top three values (8, 8, 12) with the average for those three (9, 9, 9).
- **Use random values.** You can replace real-life student IDs and school codes, for example, with randomly generated values.
- **Replace free responses.** These can unknowingly contain too much PII. You can replace them with categories. For example, when asking students how they get to school, you can use different categories for the replies (bus, on foot, car, etc.).
- **Remove instrument start and end times.** Start and end times are typically not useful for analysis. You can replace them with the duration of the activity.
- **Remove informed consent variable.** While informed consent is a critical part of data collection, you do not need to include a variable indicating whether consent was obtained for each participant.

(PII, CONT.)

strategies. If you decide to use these, please document your efforts and include this documentation in your submission, either in the Risk-Utility Assessment tab or as a separate README.txt in the Data Detail tab (see the DDL User Guide resources linked on this page for instructions).[6]

Please also indicate which indirect identifiers are present in your data (the User Guide resources explain how, and include a complete list of possible direct and indirect identifiers). After you submit to the DDL, the Data Services team will evaluate the privacy risk prior to the clearance review. If the team feels the risk is too high, they will automatically take steps to de-identify your data, using some of the methods described here (see Clearance Review for more information).

6 Often, a complete description of your de-identification methods should not be made public, because the methods could be reverse-engineered. In that case, you can write up an "abbreviated" description of your methods to share publicly. Please include both the abbreviated and complete descriptions in your submission.

EXPLORE THE DDL USER GUIDE ONLINE

How to indicate the indirect identifiers in your data?

Data Services will review this information and evaluate the privacy risk. They will automatically de-identify your data if the risk is considered too high.

How to include documentation of your de-identification strategies in your submission?

You can describe these strategies in the Risk-Utility Assessment tab when you register your data. You can also upload a separate README.txt in the Data Detail tab under "Other Reference Materials."

Risk-Utility Assessment

Risk-Utility Assessment

STAGE 3: UPLOADING DATA

CREATING AN ACCOUNT

The first step to sharing your data is creating an account on the DDL. USAID partners who submit data should create a Partner Account. See the DDL User Guide resources linked below for instructions, as well as brief tutorials on how to navigate the DDL.

EXPLORE THE DDL USER GUIDE ONLINE

An introduction for new submitters, including how to create an account and navigate the DDL.

- **Introduction and Definitions**
- **Creating an Account**
- **Navigating the DDL**
- **Development Data Library**
- **Creating a DDL User Account**
- **How to Request an Editor Account**
- **Navigating the DDL**

IMPORTANT DOCUMENTATION

Documentation is a critical part of your DDL submission. Some documents, such as informed consent and a codebook, are required. **Others, like a memo explaining how to correctly apply survey weights, are needed for analysts to make valid conclusions. Explanations of your de-identification strategies and data cleaning methods are also helpful.** You can attach these different types of documentation in the Data Detail tab when you register a dataset or data asset. Please see the DDL User Guide resources linked below for instructions.

STOP

EXPLORE THE DDL USER GUIDE ONLINE

Documentation to include in your DDL submission

USAID officials can only review your data when informed consent, a codebook, and survey questionnaire (if applicable) are attached.

How to add documents to your submission?

Use the Data Detail tab to add these files.

Preparing Data for Submission

Registering Your Data

MAKING YOUR DATA DISCOVERABLE

USAID established the DDL to make it easier for partners, academic and scientific communities, and the public at large to discover data. In the DDL User Guide resources linked on the next page, you can see how your data will appear to users of the platform. **What can you do to make your data easier for DDL users to discover and use?**

Your submission's metadata are key. Metadata are "data about data." They include the title, program area, and relevant countries, as well as keywords (or tags) to help others searching the DDL Data Catalog. Every dataset and data asset in the DDL has a primer page that presents metadata. As explained in the User Guide resources, you complete the metadata when registering a new dataset or data asset. As you do so, consider how you can make it easier for others to discover your data.

For example, when naming a dataset, we recommend the name include the country or countries where the data were collected, the name of the USAID activity, and the year the data were collected (e.g., Zambia Ready to Learn 2019). If the data were collected using a well-known instrument, like EGRA, Annual Status of Education Report (ASER), or Workforce Outcomes Reporting Questionnaire (WORQ), and if the data are from a baseline, midline, or endline, that information should also be added (e.g., Zambia Ready to Learn EGRA Baseline 2019). The more descriptive the name, the easier it will be for others to discover the data and see the connections between related datasets (see Organizing Datasets and Data Assets for more information).

Keywords are another opportunity to make your data discoverable. The country or countries where the data were collected (e.g., Zambia), the name of the USAID activity (e.g., Ready to Learn), the year (e.g., 2019), and instrument (e.g., EGRA) should all be included as keywords. If the activity targeted specific populations (e.g., disabled learners or youth) the names of those populations would also be useful keywords. Keywords can be more than one word and you can add as many as you like, separated by commas. You can add keywords in the Data Detail tab when registering your data.

There are many more opportunities to add metadata during the registration process. In the User Guide resources, you can find step-by-step instructions for completing each metadata tab. Not all fields are required, but the more information you provide, the easier it will be for others to discover and use your data. To save time, you can pre-populate the metadata if your new submission is related to an existing data asset or dataset in the DDL. You can find instructions for doing this in the User Guide resources.

STAGE 1:
PRODUCING DATA

STAGE 2:
CLEANING DATA

STAGE 3:
UPLOADING DATA

GLOSSARY

(MAKING YOUR DATA DISCOVERABLE, CONT.)

GO

EXPLORE THE DDL USER GUIDE ONLINE

How do your data appear in the DDL?

See how others can use the DDL Data Catalog, homepage, or search bar to discover your data.

Navigating the DDL

Search Features

How to complete your submission's metadata?

Step-by-step instructions for filling out each tab of the metadata.

1. Certification
2. Data Overview
3. Risk-Utility Assessment
4. Registration Profile
5. Award Information
6. Data Detail
7. Data Embargo
8. Finalize Submission

Registering Your Data

Risk-Utility Assessment

How to pre-populate metadata?

You can pre-populate metadata if your submission is related to an existing dataset or data asset.

Metadata Prepopulation

DDL ACCESS LEVELS

When registering a dataset or data asset, you have the opportunity to recommend an access level in the Risk-Utility Assessment metadata tab. There are three options:

- **Public:** The data can be made available to anyone without restrictions.
- **Restricted Public:** The data can be made available under certain restrictions. Researchers from institutions with an IRB can request access to the data using an online form (you must first create an account in the DDL to access the form). The researcher must explain the reason for requesting the data and include documentation that the institution's IRB approves the research.
- **Non-Public:** The data cannot be made available, except for internal use by the federal government.

When choosing an appropriate access level, there are a couple of things to consider. The first is informed consent. When getting consent, did you explain that the data you collect would be shared in a data repository such as the DDL? If you did not, then USAID would not be able to make your data public and the appropriate access level would be Non-Public.

A second consideration is privacy risk--the risk of identifying individuals participating in your activity. Data Services has determined that if a dataset contains both student-level information, such as student scores, and school-level information, such as school codes, the risk of identifying individuals would be too great, even if the data are masked or anonymized (see How to Treat PII for more information). The dataset would have to be published at the Restricted Public level. To publish at the Public level, you would have to "de-link" the students and schools; that is, you would have to either remove all student-level information, or all school-level information, from the dataset.

Because grouping students by school is useful for analyzing education data, we recommend keeping both student- and school-level information and selecting the Restricted Public access level. In the "Proposed Access Level Rationale" field in the Risk-Utility Assessment tab, you can select #6 (personal privacy risk). Of course, if there are other reasons for making your data restricted or even non-public, please fill out the Risk-Utility Assessment accordingly. In the DDL User Guide resources linked below, you can find a complete list of legal justifications for making data restricted or non-public.

Once you submit your data to the DDL, USAID officials will consult with your AOR/COR during the clearance review, to confirm that the access level you proposed is the most appropriate.

EXPLORE THE DDL USER GUIDE ONLINE

Selecting the right access level and rationale

More information about access levels in the DDL and legal justifications for making data restricted or non-public.

Risk-Utility Assessment

Risk-Utility Assessment

UPLOADING DATA

Ready to share your data? When it comes to adding datasets to your DDL submission, you have different options. The first is to use the "data ingest" feature to upload data directly into the platform (see the DDL User Guide resources on the next page for more information). You can also attach a dataset as a separate file in the Data Detail tab, under "Other References," although this would prevent secondary users from using the platform's data visualization features. Below, we explain some more "rules of the road" for adding datasets to the DDL.

RULES OF THE ROAD:

Wide vs. long data

There are no limits to uploading long data using the "data ingest" option. However, you cannot upload wide data with more than 500 columns. If your dataset has more than 500 columns, you will need to attach it as a separate file in the Data Detail tab.

Data in English

Data submitted to the DDL must be completely in English to be published at the Public or Restricted Public access level. If your data are not in English, please select Non-Public when choosing the access level (see DDL Access Levels for more information).

Data in a non-proprietary format

Regardless of whether you use the "data ingest" option or attach a dataset as a separate file, we recommend keeping your data in a machine-readable, non-proprietary format such as comma-separated values (CSV). Although the ingest feature will recognize Microsoft Excel and other formats, if your file has multiple tabs, the function will only upload one of the tabs. Moreover, the function will not transfer any Excel formulas that may be present in the file. If you attach your dataset as a separate file, by using non-proprietary formats, which are open-source and not owned or controlled by one company, you would make it easier for secondary users who do not have the necessary software to access your data.

Attaching a statistical-package version

We also encourage you to attach a statistical-package version, like SAS or Stata, under "Other Reference Materials." This best practice minimizes the likelihood of errors being introduced when others recreate the statistical version of your data. The statistical-package version should be identical to the data you upload using the "data ingest" option or attached in a non-proprietary format. Any efforts you make to de-identify your data should be applied to all versions of the data you submit.[7]

7 The Data Services team will not examine the statistical-package version of your data. Therefore, if the team makes any changes to your data during the Risk Assessment and Data Mitigation Review, the statistical-package version would not be made public because it would not include those changes. See Clearance Review for more information.

GLOSSARY

(UPLOADING DATA, CONT.)

EXPLORE THE DDL USER GUIDE ONLINE

How to upload your data to the DDL?

Instructions for uploading your data to the DDL using the "data ingest" feature or attaching as a separate file.

Adding Datasets (Data Ingest)

Primer Page and Dataset Creation

CLEARANCE REVIEW

What happens once you submit to the DDL? In this section, we explain the clearance review, which is a series of steps USAID takes to make as much of your data publicly available as is practical, while ensuring confidentiality and safety of Agency staff and program beneficiaries.

As shown in the graphic on the next page, the clearance review involves different USAID offices and operating units. The Data Services team will first conduct a Quality Assessment Review, which is an initial check to determine whether your submission contains the required elements, such as informed consent and a codebook. (You can find more details about the elements officials will look for in the DDL User Guide resources linked on the right.) If your submission is incomplete, Data Services will reach out to you to request additional information.

Next is the Risk Assessment and Data Mitigation Review, where officials evaluate the PII in your data and the risk of identifying individual participants in your activity. If officials feel the privacy risk is too high, they will automatically take steps to de-identify your data (see How to Treat PII for common de-identification strategies).

At the end of the Risk Assessment Review, officials will consult with the operating unit of origin (including your AOR/COR and Activity Manager) to determine whether the data can be published at either the Public or Restricted Public access level. You can learn more about how USAID determines access levels in DDL Access Levels.

If your data can be published at either the Public or Restricted Public level, the clearance review will begin. During this review, a series of USAID offices will examine your submission. These include the Privacy Office, the Office of Security, and the Freedom of Information Act (FOIA) Office.

EXPLORE THE DDL USER GUIDE ONLINE

What will USAID officials check when reviewing your DDL submission?

A list of things officials will be looking at during the initial review of your submission.

- **Submitting Your Data For Review**
- **Finalize and Submit**

USAID offices reviewing your submission

A list of the different offices that will examine your submission and the role they play.

- **Data Clearance Process**
- **Data Clearance Process**

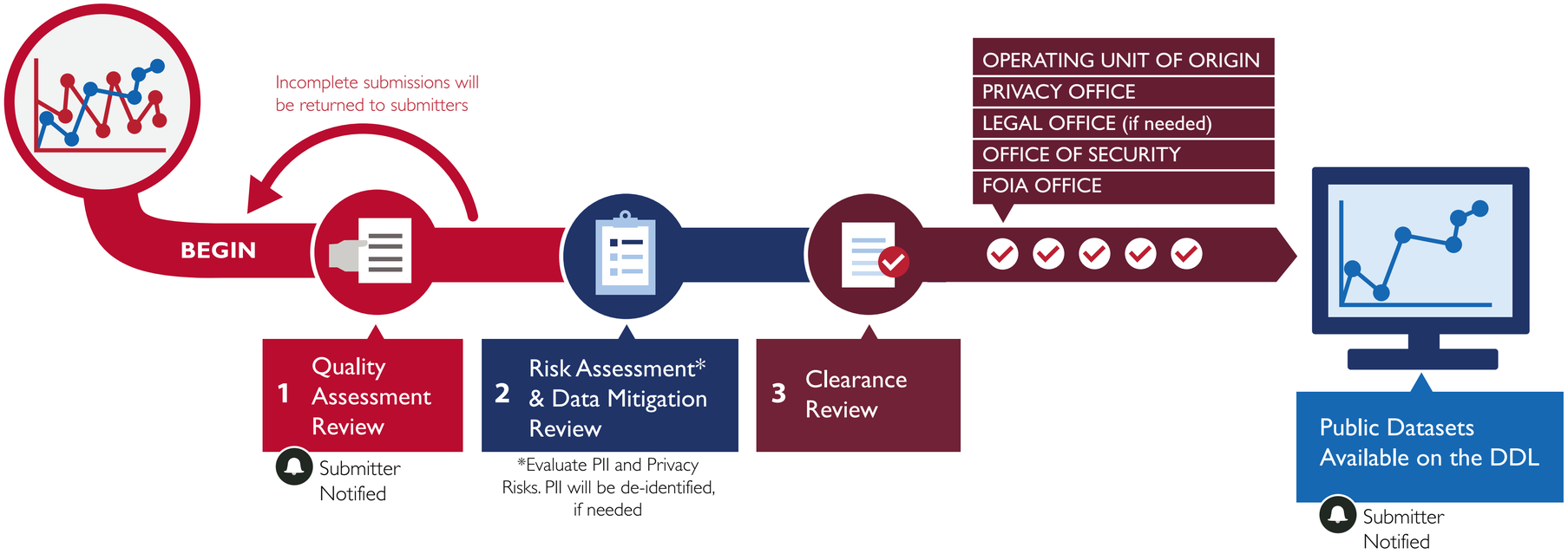
(CLEARANCE REVIEW, CONT.)

You can find more about these offices and their roles in the clearance review in the User Guide resources. Once the offices approve, your data will be published on the DDL.

If, on the other hand, officials feel it would not be appropriate to publish your data at either the Public or Restricted Public level, your submission would not go through the clearance review and the access level would be set to Non-Public. At that level, others would not be able to access your data in the DDL, although they would still be able to view the primer page and metadata.

Because every submission is different, it is hard to say how long the clearance review might last. Factors like how complete your submission is, whether the data are properly formatted, and the number of other submissions USAID officials are reviewing, all play a role in the length of the process. You will receive an email notification from USAID once your submission is under the Quality Assessment Review, and another when the clearance review is complete. You can also email the USAID Data Services team (dataservices@usaid.gov) at any time during the process if you are curious about the status of your submission.

STEPS IN THE CLEARANCE REVIEW



GLOSSARY

This glossary explains terms you may encounter while navigating the DDL, as well as those that come up more broadly when working with data.

Anonymize
To encrypt or remove personally identifiable information (PII) from a dataset so that a person cannot be identified. Names and real-life student IDs and school codes are common examples of PII in education data that should be anonymized.

Application programming interface (API)
A set of procedures that allow IT systems and databases to communicate with one another. If your data are stored on an IT system or database that is funded by USAID and has an API endpoint, you can enter the API access URL in the Data Overview tab, instead of uploading the data.

Clearance review
The process through which USAID prepares data to be shared with the public. This process can take many forms, depending on what a dataset contains and who will be able to see it when it is published.

Codebook
A spreadsheet or document that helps the end user understand what a dataset contains. The codebook defines the column headings in a dataset and the allowable values for each column, and includes any further clarification about the data in each column.

Common identifier
A variable used to merge datasets. For example, a student ID variable could be used to merge a dataset containing EGRA scores with another dataset containing student survey records.

Correspondence error
Missing or extra information in a dataset. For example, 10 students are expected per school, but only five are reported in the dataset.

CSV
A comma separated values (CSV) file is a way to store data digitally. In a CSV file, each line corresponds to a row in a table. Fields within a line are separated by commas, and each field corresponds to a column in the table. Datasets submitted to the DDL are required to be in a non-proprietary format like .CSV (although partners are also encouraged to submit versions of the data in formats produced by proprietary statistical software).

Data
Facts and statistics that are collected together and used for reference or analysis. Test scores, enrollment, teacher surveys, and student demographics are all examples of education data.

Data asset
A group of datasets with common characteristics, such as datasets that were all generated by the same survey or research study. A data asset could contain an EGRA baseline, midline, and endline, for example. Alternatively, a data asset could contain an EGRA, SSME, student survey, and teacher survey.

Database
An organized collection of data, data assets, and datasets that are usually stored and accessed electronically, and can easily be accessed, manipulated, and updated.

Data collection plan
A document that describes in detail the process of gathering quantitative and qualitative data used for monitoring, evaluation, and learning.

Data management plan (DMP)
A document that describes the data you intend to collect and how you will store, analyze, and share those data. Crafting a DMP with your AOR/COR, as part of the Monitoring, Evaluation, and Learning Plan is one of the

first steps you should take when planning an activity.

Data repository
A collection of data actively managed by a data curation team. This team uses best practices in data preservation to ensure that the data contained in the repository are accurate, well documented, and can be reliably accessed over a long period of time.

Dataset
A single data spreadsheet arrayed as a table, with a single tab and no formatting other than column headings.

De-identification
A method for removing or masking PII from a dataset before submitting to the DDL. This is essential for minimizing the privacy risk associated with publishing survey or research data. Common strategies for de-identification include using broader categories, top-coding or bottom-coding (replacing the highest or lowest values with averages), and replacing data with random values.

Development Data Library (DDL)
USAID's repository of USAID-funded, machine-readable data created or collected by the Agency and its implementing partners.

Direct identifier
Data, such as name and real-life student ID, that can be used to identify a person without additional information. If not properly handled, they may seriously compromise the privacy, security, and confidentiality of individuals whose records appear in a dataset. USAID requires the removal of all direct identifiers before data are submitted to the DDL.

Indirect identifier
Data that, standing alone, do not identify a specific person, but can be used to identify someone when combined with other information. Age, sex, and geographic subdivision are examples of indirect identifiers. Indirect identifiers can convey important information for research and removing them may reduce the usefulness of a dataset.

Informed consent
A process where potential research participants are given information about the purpose and components of the research, and are able to make an informed decision about whether to participate. Informed consent applies to participants 18 years old or older; informed assent applies to participants 17 years old or younger. Partners must include a copy of the language used to obtain informed consent or assent when submitting data to the DDL.

Instrument
The tool used to collect data, such as an assessment, questionnaire, or survey.

Long data
A dataset structured so that there are more observations than variable columns.

Machine-readable format
Refers to information or data in a format that can be easily processed by a computer. A common machine-readable format is .CSV.

Masking
Replacing direct identifiers with random values (like "9999" or "ABCDE") to preserve the form of the original data, while making association with individuals more difficult.

Metadata
Data that describe data. They can include title, description, author, sector, keywords, and other information that can help users discover and understand your data.

Non-proprietary format
A file format that is open-source and not owned or controlled by one company. .CSV is a non-proprietary format, for example, while .DTA, .SAV, and .XLS are proprietary formats.

Non-Public access level
Data assets or datasets shared at this level in the DDL are not available to members of the public. This includes data that are only available for internal use by the federal government.

Outlier
A data point that differs significantly from other observations.

Personally identifiable information (PII)
Information that can be used to identify a person.

Primer page
A section in the DDL, created for each data asset and dataset, that contains an overview of the metadata, and an option to preview and visualize the data.

Privacy risk
The potential that individual contributors to a dataset could be identified from the information in the dataset.

Public access level
Data assets or datasets shared at this level in the DDL can be made publicly available to all without restrictions.

Raw data
Data that have not been processed or cleaned for use.

Restricted Public access level
Data assets or datasets shared at this level in the DDL are available only under certain restrictions. One example, among many, is data that contain sufficient granularity or linkages that make it possible to re-identify individuals, even though the dataset has been stripped of direct identifiers.

Risk-Utility Assessment
A key part of DDL submissions, the Risk-Utility Assessment helps USAID determine the potential risk of sharing the data with various audiences, including the public.

Secondary user
Also known as "end user," a person reviewing or analyzing data who is not the original data producer.

Skip pattern
A feature of an assessment or survey where an item or question is skipped if it is not applicable. For example, if a student is unable to answer the first few items in an EGRA subtask, the assessment may skip to the next subtask.

Survey weighting
A process, after sampling, to make sure the characteristics of the sample reflect those of the larger population. For example, a student survey sample may be weighted to make sure the urban/rural divide resembles the divide in the population. Datasets submitted to the DDL should include a memo detailing the construction of survey weights, and how to correctly apply and use them for analysis. Furthermore, datasets should not only include the final weight variable, but also all the variables produced and used for weighting (such as the first stages weights, strata, and finite population correction variables).

Tabular format
Data presented in the form of a table with rows and columns.

Time-adjusted scores
Assessment scores that have been adjusted for time when a student attempts all items in a subtask in less than 60 seconds.

Wide data
A dataset structured so that there are more variable columns than observations. Datasets with more than 500 columns cannot be ingested into the DDL and should be included as a single-file attachment.

This publication was produced for review by the United States Agency for International Development (USAID). It was prepared by EnCompass LLC and its partner MSI, a Tetratech company, for the Data and Evidence for Education Programs (DEEP), Contract No. GS-10F-0245M. The views expressed herein do not necessarily reflect the views of USAID.