DEVELOPMENT DATA LIBRARY (DDL) FREQUENTLY ASKED QUESTIONS

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.

CONTENTS

1. CONSENT

1.1. WHO SHOULD I GET INFORMED CONSENT/ASSENT FROM WHEN COLLECTING DATA FROM STUDENTS AT A SCHOOL?

It depends on what you agree with the Ministry of Education in the country where you are operating, and on the types of questions you are going to be asking. Generally, it is the director or head teacher who gives you consent to come into the school to administer a survey, and you can also ask the children if they agree to participate (this is known as informed assent, since children are not legally able to give consent). However, you should check with local laws and the Ministry of Education to determine if you can get consent from the director or teachers in lieu of parental consent. In some localities, a director or teacher can act as a guardian in some situations, but you would need to make sure of that.

It also depends on the types of questions you are going to ask, and whether you are collecting information outside the scope of normal reading and math questions. Are you going to be asking questions about the children's home environment or families? About income, facilities, or assets at home? What happens at home, what languages they speak at home, or where their parents are from? If so, we recommend you get parental consent because now you are talking about questions that are not necessarily related to normal education activities--you are expanding the scope of what the assessment is. Consent from the school director or a teacher is not going to be enough because they cannot give you consent to collect information about the child's family. Getting consent from parents may require advance planning--you may need to send the consent forms out to homes in advance, prior to the school visit to administer the assessment.

When you present your questionnaire and consent procedures to an Institutional Review Board (IRB) or Ethics Review Committee (ERC), they will also be providing guidance. If an IRB or ERC determines that for your study, certain forms of consent are either required or can be waived, you should submit that documentation to the DDL with the dataset as well.

1.2. IF CONSENT WAS OBTAINED OVER THE PHONE, SHOULD I SUBMIT THE SCRIPT AS DOCUMENTATION TO THE DDL?

Yes, if you used verbal consent, you should include the script that was used to obtain consent when submitting to the DDL. See the Roadmap section "Getting Consent" and the DDL User Guide article Registering Your Data for instructions on how to include that documentation in your submission.

1.3. IF THE CONSENT STATEMENT IS GIVEN IN OTHER LANGUAGES, DOES THE DDL REQUIRE ENGLISH TRANSLATION?

Yes, we do ask that the consent statement be provided in English. Rather than backtranslate a statement from a local language into English, you can submit the original English template if it is available.

2. CODEBOOKS AND DOCUMENTATION

2.1. WHAT FORMAT IS REQUIRED FOR DATA LABELS OR CODEBOOKS?

Comma-separated values (CSV) is our preferred format for tabular information, including codebooks. The Roadmap section "Creating a Codebook" and DDL User Guide article Preparing Your Data For Submission describe best practices for documenting data in a codebook.

2.2. HOW MUCH DETAIL SHOULD BE PROVIDED IN THE CODEBOOK?

The codebook should define all the variables that are included in the dataset or datasets, as well as the type of variables included (numeric, binary, textual, etc.). If there are any sort of abbreviations used in your variable names, those should be defined as well, as well as any units of measurement that are used for a specific response to a question. If your responses have been coded categorically, we would want to know what group of answers those responses associate with. Essentially, we want all of the information about the variables that would make them usable by someone in the future, who may not have access to the team that originally collected the data.

2.3. WHEN WE SUBMIT SUPPORTING DOCUMENTATION (E.G. CONSENT STATEMENTS, QUESTIONNAIRES), WILL ALL THE SUPPORTING DOCUMENTATION BE PUBLISHED PUBLICLY IN THE DDL ALONGSIDE THE DATA?

Yes, USAID would publish that supporting information along with the data record. However, they may publish a different version from what was submitted, depending on whether the data are modified to make public.

3. DATA MANAGEMENT & ORGANIZATION

3.1. CAN YOU CLARIFY THE DIFFERENCE BETWEEN A DATASET AND A DATA ASSET?

A data asset can be thought of as a "shell" that contains datasets within it. When uploading data to the DDL, you first need to create the data asset, even if you will only be submitting a single dataset. A data asset contains datasets that are related to one another--for example, a baseline and a midline, or student assessment scores, student interviews, and teacher interviews. To save time, you can pre-populate the metadata of a new dataset if it is related to a data asset or dataset that you have already created. You can find instructions for doing so in the DDL User Guide video Metadata Prepopulation. The Roadmap section "Organizing Datasets and Data Assets" also describes in more detail the relationship between datasets and data assets.

3.2. PLEASE PROVIDE CLARIFICATION OR EXAMPLES REGARDING WHAT IS CONSIDERED UNSTRUCTURED VS. STRUCTURED QUALITATIVE DATA.

Unstructured qualitative data might simply be a long string of text, such as interview questions and answers in a Microsoft Word document. Structured qualitative data would be data that have been changed into categorical data in some way or categorized during the analysis process. That is what we prefer you submit to the DDL. For an example of structured qualitative data that have been submitted to the DDL (at the Public access level), please see the Pacific Islands Global Climate Change Performance Evaluation: Qualitative Data.

3.3. CAN I CREATE A DATA ASSET THAT CONTAINS BOTH QUALITATIVE AND QUANTITATIVE DATA? OR SHOULD I UPLOAD THEM AS SEPARATE SUBMISSIONS?

In general, yes, you could use a single data asset and have several datasets within it, including both qualitative and quantitative data. USAID ADS 579, which directs what should be submitted to the DDL, specifies that qualitative data can be included, although such data should be structured. In addition, qualitative data are often treated differently in the DDL clearance process because such data can sometimes reveal a lot of PII and might be inappropriate for broader distribution. In general, however, it would make sense to keep that conceptual connection between the two pieces of the data asset.

3.4. I AM INTERESTED IN THE BASIC REQUIREMENT TO INCLUDE A DATA MANAGEMENT PLAN, WHICH WAS ALSO REFERENCED IN THE ADS 579. CAN YOU PLEASE SPEAK TO THAT?

There is some initial guidance in the Roadmap in terms of what should be included in a data management plan for USAID activities (see “Data Management Plan”), and the Agency is currently developing more guidance on this. There is a working group within USAID that is working to ensure that data management plan requirements reflect different needs of different functional units. This group is under the auspices of the DATA Board, which is the Agency's data governance body that was formed late last year.

4. NAVIGATING THE DDL

4.1. HOW CAN I ACCESS AND USE DATA IN THE DDL? I'VE TRIED TO SEARCH FOR DATASETS TO USE, WITHOUT SUCCESS.

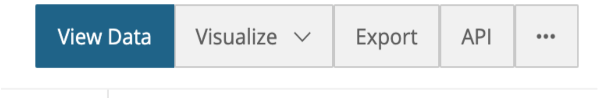
The DDL User Guide article Navigating the DDL describes how to search the catalog. There are many filters available in the catalog to browse data in the DDL related to a particular country, sector, or initiative, to name a few of the possibilities.

4.2. I CREATED MY ACCOUNT TO ACCESS THE DDL, BUT WHEN I TRY TO ACCESS PUBLIC DATASETS, I ALWAYS GET THE MESSAGE, "YOU DO NOT HAVE PERMISSION TO VIEW THIS DATASET" OR "THE CURRENT USER DOESN'T HAVE ACCESS TO THIS RESOURCE." HOW CAN I ACCESS PUBLIC DATASETS?

If you look at the record for any dataset on the DDL, you will see the field “Proposed access level,” and this will say either Public, Restricted Public or Non-Public. Public data should be available--it may be a matter of just clicking on the associated record for the data set, but an ingested data set should appear right there on your screen. If the data are restricted, USAID has a request form that you can fill out to ask for access to the data. The form includes language and assurances that the data will be used by researchers who can ensure the safety of the data and that the data will not be redistributed. Non-Public data, however, are not available for request. USAID is still obligated to present a record showing that the Agency owns that data, but for a variety of reasons related to privacy and security, they are simply not able to share the dataset itself. So, look for that proposed access level field. For restricted data, Data Services puts a note into the proposed access level comment field that includes the link to the request form.

4.3. WE ARE OFTEN INTERESTED IN USING PUBLIC DDL DATA AS EXTERNAL DATA SOURCES OR AS ADDITIONAL DATA LAYERS TO COMBINE/COMPARE WITH OUR DATA. WOULD WE BE ABLE TO ACCESS THE DDL PUBLIC DATA SOURCES DIRECTLY THROUGH A LINK, AN API, OR BY CONNECTING TO THEM USING TOOLS LIKE POWER BI/TABLEAU?

Each dataset automatically receives an API endpoint. Just click on the API link for the dataset (as seen in the image below) to receive the API endpoint and links to relevant documentation.



4.4. HOW CAN I MAKE IT EASIER FOR DDL USERS TO DISCOVER MY DATA?

The Roadmap section "Making Your Data Discoverable" and the DDL User Guide article Preparing Your Data For Submission describes best practices for naming and describing data to make them easier to find.

4.5. HOW CAN WE STAY UP-TO-DATE ON INFORMATION ABOUT THE DDL?

The Data Services team produces a monthly newsletter called "Insights," which often has information about the DDL in it. You can complete this form to ask to be added to the subscription list. If you want to know the status of a particular submission you made, you can email Data Services at dataservices@USAID.gov.

5. PERSONALLY IDENTIFIABLE INFORMATION (PII)

5.1. WHAT PERSONAL IDENTIFICATION SHOULD BE EXCLUDED FROM SUBMISSION?

You should not include direct identifiers (such as names or real-life IDs) that could directly identify an individual without any other information. See the Roadmap section "How to Treat PII" for more information.

5.2. I WAS UNDER THE IMPRESSION THAT EGRA DATA WERE TREATED AS A PRIVACY SUBMISSION. DOES USAID NOW ENCOURAGE THESE DATA TO BE PUBLIC?

Where possible and where context allows, the Office of Education is encouraging the submission of EGRA data to be available for public consumption. We recognize, though, that EGRA data often contains information on human subjects that cannot be made widely public. For that reason, we generally expect datasets to come in at the Restricted Public access level, which would require someone to apply to access the data and tell us what they are using the data for. We recognize the value of EGRA data for the wider education sector for a variety of purposes. We also recognize that depending on the context, the EGRA data that were collected may not be suitable for even Restricted Public access. But either way, we encourage education data producers to submit the data to the DDL, and through that process tell us what you think in terms of the proposed access level. We really want to make sure that we are getting as much data as possible to create a solid body of education evidence.

5.3. WOULD THE NAME OF THE SCHOOL BE A DIRECT OR INDIRECT IDENTIFIER, OR WOULD IT DEPEND ON THE CONTEXT OF THE SURVEY?

The name of the school is considered an indirect identifier, because it could be used to help identify individuals but would not, on its own, directly identify an individual. You do not necessarily have to remove the name of the school before submitting to the DDL, but just know that Data Services will not publish the name of the schools. We do not discourage you from submitting this information because we do want you to give us really granular datasets, because we want to have those available for follow-on activities. In the event that USAID needs to conduct its own study, or five or ten years down the road it needs to compare another assessment with this assessment, we want the data to be as granular as we can possibly get it. However, although we collect really granular data, we do not publish the really granular data.

5.4. IS THERE EXPLICIT GUIDANCE ON WHICH GEOGRAPHIC SUBDIVISIONS NEED TO BE STRIPPED, OR IS THIS A SUBJECTIVE DECISION?

What we are trying to do when you submit data to the DDL is make sure that we are not able to directly identify the individuals who participated in the survey. However, we do want the data to be as useful as possible to the Agency so that USAID can conduct follow-on activities. So you can submit that really low-level, granular geographical data. When Data Services is getting ready to publish such data, however, they would likely roll it up to the district level or something similar before making the data publicly available. But when we are talking about internal use at USAID, we want the data to be as granular as we can possibly get it.

5.5. ARE WE THEREFORE CREATING DIFFERENT DATASETS (FOR USAID, HOST COUNTRY GOVERNMENTS, RESEARCHERS, ETC.)?

Possibly, because you are probably sharing really granular data with host country governments and with USAID, but what we share with the public or with research institutions is going to be a different version of the dataset. The data you are sharing with USAID should not be very different from what you might share with a host country government. If you are sharing direct identifiers like student names, that might be the difference between what you share with USAID and what you share with the Ministry of Education.

6. SUBMITTING DATA

6.1. WHAT KIND OF DATA ARE SUBMITTED TO DDL?

All kinds of data are submitted to the DDL--any data that result from Agency-funded projects. This could range from supply-chain data for HIV resources to surveys of small farming efforts under the Feed the Future initiative, to data from the Office of Human Capital and Talent Management that looks at hiring trends among USAID staff. USAID asks that essentially all microdata be submitted. The question of publication level (Public, Restricted Public, and Non-Public) will emerge from the submission form and from USAID's review of the data once they are received. But having the data submitted to the Agency in the first place is key for long-term preservation, access, and reuse.

6.2. CAN USAID PROVIDE MORE GUIDANCE ON WHAT KIND OF DATA FROM AN EDUCATION PROGRAM WOULD BE REQUIRED FOR SUBMISSION, ASIDE FROM EVALUATION DATA?

Under ADS 579, any data used to support the creation of an intellectual work should be submitted to the DDL. Typically, this is research/evaluation data and survey data meant to inform program design--you can find a more detailed list in ADS 579. It does differ quite a bit by activity, so we recommend consulting with an AOR/COR on the specifics for any activity.

6.3. FOR TESTS AND ASSESSMENT, DO ITEM LEVEL DATA NEED TO BE SUBMITTED? ARE AGGREGATED SCORES ENOUGH?

Item level data are preferred as they are more useful for analysis.

6.4. IN WHAT FORMATS WILL THE SYSTEM UPLOAD DATA?

We recommend keeping your data in a machine-readable, non-proprietary format such as CSV. However, the "data ingest" function does recognize some other formats, including: Tab-separated values (TSV), Microsoft Excel (XLS), Microsoft Excel (OpenXML), ZIP archive (shapefile), JSON format (GeoJSON), Keyhole Markup Language (KML), and Zipped Keyhole Markup Language (KMZ).

6.5. CAN DATASETS WITH OVER 500 VARIABLES BE SUBMITTED?

They certainly can be submitted. Unfortunately, there is a limit to the "data ingest" function of the platform. This function allows the platform to read the data file and make it available for visualization and other kinds of analysis within the platform. If your dataset has more than 500 variables, the platform will not be able to ingest it for those in-platform uses, but you can still include it as a separate attachment in the data detail tab of the submission form. So we would just ask that you attach the dataset if it's too big to ingest. See the Roadmap section "Uploading Data" for more information.

6.6. CAN AN EXCEL FILE BE SUBMITTED TO THE DDL?

We encourage you not to submit Excel files for a couple reasons. First, the "data ingest" function will not recognize any Excel formulas that may be present in your data. And second, if your Excel file has multiple tabs, only one of those tabs will be uploaded into the platform. We encourage you instead to use CSV files both for uploading data using the "data ingest" function as well as attaching as a separate file. CSV is also a non-proprietary format that is great for preservation purposes, where we do not need to worry about versioning over time and different features that may have been common in past versions but that will not be in the future.

6.7. HOW CAN I DEAL WITH DATA WITH MIXED NUMBERS OF ROWS?

Sometimes we see submissions where multiple tables are included in a single spreadsheet. If this is what the question is referring to, the best practice is to put each of these tables in a separate CSV file, so they can each be ingested by the platform.

6.8. CAN FOREIGN LANGUAGE DATA BE UPLOADED TO THE DDL?

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.

6.9. IF WE HAVE NOT YET SUBMITTED OLDER DATA TO THE DDL, SHOULD WE BACK-DATE WHEN SUBMITTING?

That will depend on the terms of the award or contract under which the data were collected. You may have awards that required submission where you have not yet submitted the data. That would be something to discuss with the AOR/COR. In general, the guidance provided in the Roadmap is intended for data submissions you will make in the future.

7. TIMEFRAME FOR SUBMISSIONS

7.1. AT WHAT POINT AFTER A DATASET IS FIRST USED TO PRODUCE AN INTELLECTUAL WORK SHOULD THE DATA BE UPLOADED TO THE DDL?

Generally, the terms of the awards and contract state when the data should be submitted. Data Services frequently sees an influx of submissions as the fiscal year ends and as contracts and awards start to wrap up. The sooner partners can submit their data, the better, because the fresher the data are in the team's mind, the more clearly they will be able to answer the questions on the submission form and speak with our curation team if there are follow-up questions.

7.2. HOW LONG DOES IT TAKE TO PUBLISH DATA ON THE DDL?

This can depend on a few things, including how many submissions happen to be coming in and the workload of the various teams that review the submission. These teams include the curation team as well as the risk assessment team, which does a fine-grain analysis of the variables contained in a dataset, looking for direct and indirect identifiers. In addition, the clearance process runs through four different offices at the agency. Generally, it can take several months to go from submission to publication of the record on the DDL. See the Roadmap section “Clearance Review” for more information.

7.3. HOW FREQUENTLY IS THE SYSTEM UPDATED IN TERMS OF NEW DATA SUBMISSIONS?

We get new submissions all the time and publish new items that have completed the clearance process on a weekly basis.