Dataset Requirements for Compliance with Gates Foundation Open Data Policy

All Alive & Thrive (A&T) research funded by the Bill & Melinda Gates Foundation (BMGF) must comply with BMGF data open access requirements. All A&T contractors must submit datasets from A&T research to A&T in an open access-suitable format. Note that this policy applies to quantitative datasets only.

Activity
Deliver quantitative datasets, metadata, and key associated documents in a form that allows Alive & Thrive to make them publicly available.

Deliverables and specifications

1) **Raw, cleaned data**
    1. Anonymized and all sensitive and personally identifiable information removed. This includes all names, locations, birthdates etc. In some situations, small sample sizes in geographic locations, locations in which very few households reside or unique combinations of responses may present disclosure risks. These circumstances will need to be assessed on a case-by-case basis to determine how to protect anonymity.
    2. Formatted to correspond with codebooks and questionnaires. Module order and variables should correspond with questionnaires and codebooks.
    3. Variables and their values labeled exactly as they appear in questionnaires, with clear and concise naming conventions so that users can easily link dataset values, codebooks, and questionnaires. A recommended best practice is to write format/reshape, label, cleaning, anonymization files separately for each questionnaire.

2) **Codebooks**
    Provide a document detailing the rules and definitions used for coding the data. This is useful when open-ended responses are coded into quantitative data and the codes are not provided on the original data collection instrument. For each variable, the following information should be provided:

    1. The exact question wording or the exact meaning of the datum. Sources should be cited for questions drawn from previous surveys or published work.
    2. The text of the question integrated into the variable text. If this is not possible, it is useful to have the item or questionnaire number (e.g., Question 3a), so that the archive can make the necessary linkages.
    3. Universe information, i.e., who was actually asked the question. Documentation should indicate exactly who was asked and was not asked the question. If a filter or skip pattern indicates that data on the variable were not obtained for all respondents, that information should appear together with other documentation for that variable.
    4. Exact meaning of codes. The documentation should show the interpretation of the codes assigned to each variable. For some variables, such as occupation or industry, this information might appear in an appendix.

5. Missing data codes. Codes assigned to represent data that are missing. Such codes typically fall outside of the range of valid values. Different types of missing data should have distinct codes.
6. Unweighted frequency distribution or summary statistics. These distributions should show both valid and missing cases.
7. Imputation and editing information. Documentation should identify data that have been estimated or extensively edited.
8. Details on constructed and weight variables. Datasets often include variables constructed using other variables. Documentation should include "audit trails" for such variables, indicating exactly how they were constructed, what decisions were made about imputations, and the like. Ideally, documentation would include the exact programming statements used to construct such variables. Detailed information on the construction of weights should also be provided.
9. Location in the data file. For raw data files, documentation should provide the field or column location and the record number (if there is more than one record per case). If a dataset is in a software-specific system format, location is not important, but the order of the variables is. Ordinarily, the order of variables in the documentation will be the same as in the file; if not, the position of the variable within the file must be indicated.
10. Variable groupings. Particularly for large datasets, it is useful to categorize variables into conceptual groupings.
11. If variable names and variable labels contain abbreviations. These should be standardized and described and included in the codebook.

3) Data collection tools (questionnaires or surveys), tools flowchart, and study protocol

1. Questionnaires in each language administered. Original tools in English or if tools were translated into English, these should also be included, even if data collection was completed in local language other than English.
2. Other data collection instruments or tools such as registries, exams/tests, screening forms, observation checklists, and all other.
3. Descriptions of the circumstances in which each instrument or tool was used should also be provided.
4. Guides for using the data collection tools (questionnaires or surveys), these are the same guides used for training the team who collected data, including full detail on how interviews were administered, including probes, interviewer specifications, use of visual aids and all other files used during training of data collectors should be included.
5. Flowchart of the data collection tools. A graphical guide to the data, showing which respondents were asked which questions and how various items link to each other. This is particularly useful for complex questionnaires or when no hardcopy questionnaire is available.
6. Study Protocol

4) Descriptive Metadata, Reports, Publications

1. **Principal investigator(s):** Principal investigator name(s) and affiliation(s) at the time of data collection.
2. **Title:** Official title of the data collection.

3. **Contact.**
4. **Funding sources:** Names of funders, including grant numbers and related acknowledgments.
5. **Data collector/producer:** Persons or organizations responsible for data collection, and the date and location of data production.
6. **Project description:** A description of the project, its intellectual goals, and how the data articulate with related datasets. Publications providing essential information about the project should be cited. A brief project history detailing any major difficulties faced or decisions made in the course of the project is useful.
7. **Date**: Date dataset was created.
8. **Keyword(s):** Key terms that describe important aspects of the dataset.
9. **Substantive, temporal, and graphic coverage of the data collection:** Descriptions of topics covered, time period, and location.
10. **Data source(s):** If the dataset draws on resources other than surveys, citations to the original sources or documents from which data were obtained.
11. **Unit(s) of analysis/observation:** A description of who or what is being studied.
12. **Sample and sampling procedures.** A description of the target population investigated and the methods used to sample it (assuming the entire population is not studied). The discussion of the sampling procedure should indicate whether standard errors based on simple random sampling are appropriate, or if more complex methods are required. If weights were created, they should be described. If available, a copy of the original sampling plan should be included as an appendix.
13. **Final Project Report.**
14. **Appropriate Summary Statistics.**
15. **Related Publications:** Citations to publications based on the data, by the principal investigators or others.
16. **Citation**: The full bibliographic citation(s) for related publications.
17. **URLs to related publications. Technical information on files:** Information on file formats, file linking, and similar information.