B11: A Collection of Global Guides and Tools on Data Cleaning

- Who is this tool for? Health and other social protection practitioners looking to develop their processes and procedures for data cleaning, in preparation for a multi-agency population targeting data linkage initiative.
- How was it produced? Data cleaning and records matching was a key barrier to one of the implementation case countries in this collaborative, and as such a problem solving workshop was held in July 2021 to discuss the steps and stages that needed to be undertaken. This document summarizes the key discussion points and global evidence/best practice presented.

Data cleaning is a vital step which needs to be done to facilitate sharing of data across agencies for targeting purposes. Data cleaning (also known as data cleansing) is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. It refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the "dirty" data. The inconsistencies may have been originally caused by user entry errors, by corruption in transmission or storage of data, or by different data definitions. Data cleaning differs from initial data validation in that validation is performed at the time of entry, rather than on batches of data after the original entry process.

When preparing to share data across agencies for targeting purposes, ideally each agency will have done an agency-specific data cleaning on the data which is to be shared with other agencies in order to minimize additional challenges in the later data matching process which are due only to dirty data. In preparing for a data cleaning exercise, there are three big picture questions to ask first:

1. How "dirty" are the data which need cleaning? You need to realistically forecast how much "dirt" you expect to find in your data before setting a "cleaning budget" and estimating the time that will be needed for the cleaning process. For example, if the source data is very old, that will increase the share of inaccurate records, even where the data were originally quite clean. Similarly, if conventions for recording names in one part of the country, or for specific population groups, are different to the typical conventions, that would increase the likelihood of dirty data when cleaning.
2. Can the data be transformed using standard tables, such as locality tables, mapping tables, postal-code tables, facility tables, etc ? This plays two roles – helping to standardize while cleaning the data. This will also accelerate later data matching with other agencies if common data standards and definitions are used.
3. When will you stop? That is, what is your goal? This is a crucial question which needs to be known before starting the data cleaning process. How clean is "good enough" for the cleaning exercise (90%, 95%, 99%) ? As a rule of thumb, if the goal is less than 90% clean data after the cleaning process, it probably not worth doing, as the transformed DB will never be "authoritative".

There is a common set of data cleaning processes which are carried out, and a number of them have easily available software to assist in the process. The steps are:

1. **Data auditing** – this involves checks for anomalies and contradictions in the data against a set of specific constraints (e.g., on age or gender or location) and uses code to check data for violation of the constraints. Multiple software exists to carry out this step but common programs such as Microsoft Access and File MakerPro do simple constraint-by-constraint checks with little or no programming required. Some common constraints against which data can be checked are:
    - *Data-Type Constraints*: values in a particular column must be of a particular datatype, e.g., numeric, date, boolean etc.
    - *Range Constraints*: typically, numbers or dates should fall within a certain range.
    - *Mandatory Constraints*: certain columns cannot be empty.
    - *Unique Constraints*: a field, or a combination of fields, must be unique across a dataset.
    - *Set-Membership constraints*: values of a column come from a set of discrete values, e.g,, a person's gender may be male or female (or other if legally allowed).
    - *Foreign-key constraints*: as in relational databases, a foreign key column cannot have a value that does not exist in the referenced primary key.
    - *Regular expression patterns*: text fields that have to be in a certain pattern. For example, phone numbers may be required to have the pattern (999) 999–9999.
    - *Cross-field validation*: certain conditions that span across multiple fields must hold. For example, a patient's date of discharge from the hospital cannot be earlier than the date of admission.
    - For a full list, see *The Ultimate Guide to Data Cleaning. Omar Elgabry https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4*
2. **Parsing**: for the detection of syntax errors in the data. Data parsing is the process of taking data in one format and transforming it to another format (e.g., from HTML format into some other data format). It involves application of user-specified business rules in order to understand and validate any type of data *en masse*, and, if required, improve its structure in order to make it fit for purpose. A parser decides whether a string of data is acceptable within the allowed data specification. There are common software available for parsing (https://www.scrapingbee.com/blog/data-parsing)
3. **Duplicate elimination**: Eliminating duplicate data is one of the most important elements of data cleaning. Duplicate detection requires an algorithm for determining whether data contains duplicates. Usually, data is sorted by a key that would bring duplicate entries closer together for faster identification. There are multiple approaches and tools for removing duplicates: blocking and windowing; software such as DYNSI, PSNM, Dedup, DCS++. Or simple data rules can be applied, e.g., in Philippines Listahanan, a simple algorithm is used to extract all with same 1st 3 letters of last name, 2nd 2 letters of first name then middle initial and birthday.
4. **Statistical methods**: There are different statistical methods which can also help to identify problematic data in the cleaning process. By analyzing data using the values of mean, standard deviation, range, or clustering algorithms, it is possible to find values that are unexpected and probably incorrect. Although correction of such data is difficult since the true value is not known, it can be resolved by setting the values to an average or other statistical value. Statistical methods can also be used to handle missing values which can be replaced by one or more

plausible values, which are usually obtained by data augmentation algorithms. These follow agreed rules for dealing with missing values.

5. **Data transformation**: Data transformation allows the mapping of the data from its given format into the format expected by the other DB or application with/through which data will be shared. This includes for example value conversions or translation functions, as well as normalizing numeric values to conform to minimum and maximum values. This is ideally done at source within the agency doing the data cleaning but can also be applied by the receiving agency software or middleware.

There is often **a sequence in data cleaning processes**. A first step would be removing duplicates and/or irrelevant observations (e.g., data on non-citizens who may not be eligible for a program). A second is fixing structural errors in the data such as incorrect naming conventions, typos, upper/lower case conventions – things that do not conform to the data dictionary. Third would be filtering unwanted outliers which may be due to mistakes (people who are 130 years old for example). Fourth would be having a process for handling missing data (do you drop those observations, or impute values based on other observations, or handle null values in data processing?). The final step is data validation and quality checks. *https://www.tableau.com/learn/articles/what-is-data-cleaning*.

Further resources:

A useful checklist for data cleaning exercises is provided in one of the annexes of ACAPS, 2016. Data Cleaning. The whole paper is also a useful guide to data cleaning.
https://www.acaps.org/sites/acaps/files/resources/files/acaps_technical_brief_data_cleaning_april_2016_0.pdf

Another useful resource which covers data management more broadly but has a section on data cleaning is Centre for Disease Control. Data Management participant workbook, 2013.

https://www.cdc.gov/globalhealth/healthprotection/fetp/training_modules/10/managing-data_pw_final_09252013.pdf

If using EXCEL software, the following may be useful: Top ten ways to clean your data. Excel for Microsoft 365 Excel 2021 Excel 2019 Excel 2016 Excel 2013 Excel 2010 Excel 2007

https://support.microsoft.com/en-us/office/top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19