**DTM Data Quality Management**

(Data Validation)

**Executive Summary**

**Background**

**Scope**

**Features**

**What is Data Editing?**

Data editing is the activity aimed at detecting and correcting errors in data. The editing procedure usually includes three phases i.e. defining rules, data validation or data checking and correction of the errors.

**What is Data Validity?**

An activity aimed at verifying whether the value of a data item comes from the given set of acceptable values.

Data validation assesses the plausibility of data: a positive outcome will not guarantee that the data is correct, but a negative outcome will guarantee that the data is incorrect.

**Data validation procedure**

Data validation is a decisional procedure ending with an acceptance or refusal of data as acceptable. The decisional procedure is generally based on rules expressing the acceptable combinations of values. Rules are applied to data. If data satisfy the rules, data are considered valid for the final use they are intended to.

The data validation process is an iterative procedure based on the tuning of rules that will converge to a set of rules that are considered the minimal set of relations that must be necessarily satisfied.

When the validation fails, it may produce three types of error (the severity):

* Fatal error: the data are rejected;
* Warning: the data can be accepted, with some corrections or explanations from the data provider;
* Information: the data are accepted.

**Why data validation - Relationship between validation and quality**

The purpose of data validation is to ensure a certain level of quality of the final data.

Nevertheless, quality has different dimensions in official statistics: relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability, coherence, completeness. Hence, it is important to establish on which components data validation is concerned with.

Data validation focuses on the quality dimensions related to the ‘structure and content of the data’, that are accuracy, comparability, coherence.

**Accuracy**

Data accuracy refers to whether data values are correct. To be correct a data value must be both the right value and be represented in an unambiguous form, two characteristics of data accuracy are form and content.

**Form**

Form is important because it eliminates ambiguities about the content. Form dictates how a data value is represented, For example P-Codes of a country (Yemen) could be recorded as ‘0021’ but to remove the ambiguity the standard P-Code should be used with country name prefix i.e. YE in this case to make P-Code ‘YE0021’. Example from Yemen:

<https://displacement.iom.int/datasets/yemen-returnees-dataset-round-32>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assessed District | District PCode | Location Name | Location Name A | Location PCode |
| Al Wade'a | 1208 | Al-Wadhee'e | Al-Wadhee'e-الوضيع | 1208210012 |
| Al Wade'a | 1208 | Amkado | Amkado-امكدو | 1208211302 |

**Content**

As for content, “two data values can be both correct and unambiguous yet still cause problems.” This is a common challenge with free-form text, such as a city name. “The data values Nyanza-Lac and Nyanza Lac may both refer to the same city, but the recordings are inconsistent, and thus at least one of them is inaccurate.” Consistency is a part of accuracy, because “inconsistent values cannot be accurately aggregated and compared. Since much of data usage involves comparisons and aggregations, inconsistencies create an opportunity for the inaccurate usage of data.”

<https://displacement.iom.int/datasets/burundi-baseline-assessment-round-33>

<https://displacement.iom.int/datasets/burundi-baseline-assessment-round-35>

Burundi Round 35 Burundi BL 33

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Admin1 | Admin2 | Pcode |  | Admin1 | Admin2 | Pcode |
| Rumonge | Bugarama | BDI018011 |  | Bujumbura Rural | Mutimbuzi | BDI002010 |
| Makamba | Nyanza-Lac | #N/A |  | Makamba | Nyanza Lac | BDI010005 |

**Validity versus Accuracy**

Defining all values that are valid for a data element is useful because it allows invalid values to be easily spotted and rejected from the database. However, we often mistakenly think values are accurate because they are valid. For example, if a data element is used to store the name of a city. The value of ‘123’ would be invalid but the value of ‘Juba’ would be valid but inaccurate if the data is being collected in Yemen. Another example would be Yes/No. The following example is from South Sudan Round 1

<https://displacement.iom.int/datasets/south-sudan-baseline-assessment-round-1>

|  |
| --- |
| **IDP Arrival Jan-June 2017 other reason Y/N** |
| no |
| 0 |
| 0 |
| no |

**Can 100% Data Accuracy be achieved?**

The short answer is no, You can get accurate data to a degree that makes it highly useful for all intended requirements.

**Coherence and comparability**

The general definition of coherence and comparability claims that statistics should be consistent internally, over time and comparable between regions and countries. The process of detecting errors and flagging them should be coherent and consistent internally and between countries, based on common standards with respect to the scope.

**Clarity and accessibility**

To meet the requirements for accessibility, it is often seen as sufficient to make data available via the internet, whereas clarity is seen as satisfactory if a few footnotes or links to definitions are provided. If users cannot easily access data in the format they need, or if they do not understand the associated metadata, the data have little real value, even if they are perfectly accurate and coherent.

**Accessibility**

It is therefore clear that to meet the needs of users in terms of accessibility, DTM teams should offer data in a different and useable formats i.e. XLSX, CSV, JSON APIs etc. and the shared files should be consistent like the first sheet should have raw data and there is no unnecessary sheets of data in the shared file like local analyses and/or visualizations.

**Clarity – Good Metadata**

Data are most valuable when they are easily accessible and accompanied by good metadata. Consistency of definitions is very important to understand data. It must be clearly defined who is a displaced person in a context where the data was collected. Clarity of data (metadata) helps resolve ambiguity.

**Final Remark on Validation and Quality**

A final remark about the concept that data validation “aims at verifying whether data have a certain level of quality”: It is indeed true that data validation cannot ‘ensure’ a level of quality. What it can more realistically provide is that at least a certain level of data consistency considered as the minimum requirement for having acceptable data, is reached. This results not in perfect data, but in ‘plausible’ data.

**How to perform data validation: validation levels and validation rules**

Because of the variety of validation steps and procedures, it is desirable to be able to judge to what extent a data set has been validated (validation level) by validation procedures applied to it. Moreover, as statistical processes age and mature, the number of validation procedures and rules tend to grow organically, generating a need for maintenance.

It is generally assumed that there are basically two general categories:

1. Technical integrity of the file, i.e., consistency with the expected IT structural requirements (Structural Validation)
2. Logical and statistical consistency of the data (Content Validation)

The second category is generally split into different sub-categories (levels) involving more and more information. The two general categories can then be expanded forming the following validation levels from a business perspective.

* Validation level 0: consistency with the expected IT structural requirements
* Validation level 1: consistency within the data set
* Validation level 2: consistency with other data sets within the same domain and within the same data source
* Validation level 3: consistency within the same domain between different data sources
* Validation level 4: consistency between separate domains in the same data provider
* Validation level 5: consistency with data of other data providers

**Validation level 0: consistency with the expected IT structural requirements**

At this level, it is checked the consistency of the data with their expected IT requirements, for instance

* if the fields are from the verified list of questions (data dictionary)
* if the file has been sent/prepared by the authorized authority (data sender);
* if the column separator, the end of record symbol are correctly used (structure of the file);
* if the file has the expected number of columns (agreed format of the file);
* if the column have the expected format of the data (i.e., alphanumeric, numeric, etc.)

For these quality checks only the structure of the file or the format of the variables are necessary as input.

**Validation level 1: consistency within the data set**

It is checked the consistency within the elements of the data set. For these quality checks, it is needed only the (statistical) information included in the file itself.

For instance:

* check whether the number in column ‘Number of IDPs’ is not negative (as expected);
* check whether the year in the column ‘Interview Date’ is 2018, as in the file name;
* check whether the content of the 3rd column is one of the codes of the dictionary "Sex of KI";
* based on information available before data collection (for example from previous survey or other sources) one could establish a "plausibility range" for a certain variable (for instance number of components of a household).
* check consistency at (micro-level) of two (or more) variables: a certain combination of codes is illogical, a variable has to be reported only for a certain combination of codes.
* check consistency at macro-level of two (or more) variables: Total inhabitants = male inhabitants + female inhabitants, or Female inhabitants = (total inhabitants / 2) +/- 10%

**Validation level 2: consistency with other data sets within the same domain and within the same data source**

Validation levels 2 is concerned with the check of consistency based on the comparison of the content of the file with the content of "other files" referring to the same statistical system (or domain) and the same data source.

For instance:

* Case a) the "other files" can be other versions of exactly the same file. In this case the quality checks are meant to detect "revisions" compared to previously sent data. Detection and analysis of revisions can be useful for example to verify if revisions are consistent with outliers detected in previous quality checks (corrections) or to have an estimate of the impact of the revisions in the "to be published" results, for the benefit of the users.
* Case b) the "other files" can be versions of the same data set referring to other time periods. These checks are usually referred to as "time series checks" and are meant to verify the plausibility of the time series.
* Case c) the "other files" can refer to other data sets from the same data provider (e.g., missions), referring to the same or other correlated time periods. Sometimes a group of data sets (same country, same reference period) is sent at the same time.

Example: three files could be sent at the same time, from the same mission and referring to the same time period: one file includes data for "IDPs", one for "Returnees" and one for "total". Consistency between the results of the three files can be checked.

Another example: results from annual data sets can be compared with the results of the corresponding quarterly data sets.

**Validation level 3: consistency within the same domain between different data sources**

Validation levels 3 is concerned with the check of consistency based on the comparison of the content of the file with the content of "other files" referring to a different data provider on the same harmonized statistical system or domain (sharing common standards with respect to scope, definitions, units and classifications in the different surveys and sources).

For instance:

Case d) the "other files" can refer to the same data set, but from another data provider (e.g., Mobility Tracking data from different missions). Mirror checks are included in this class. “Mirror statistics involve coherence, geographical comparability as well as accuracy issues”. Often such statistics is important for data analysis at global level. Mirror checks verify the consistency between declarations from different sources referring to the same phenomenon, e.g., People moved from point A to B are same to the people reached at Point B from A.

**Validation level 4: consistency between separate domains in the same data provider**

Validation level 4 could be defined as plausibility or consistency checks between separate domains available in the same Institution. The availability implies a certain level of "control" over the methodologies by the concerned Institution.

These checks could be based on the plausibility of results describing the "same" phenomenon from different statistical domains. Examples: unemployment from registers and from Labour Force Survey, or inhabitation of a dwelling (from survey of owners of houses and dwellings vs. from population register)

Checks could also be made between results from correlated micro-data and macro-data sources.

Other plausibility checks could be based on known correlations between different phenomena: for example external trade and international transport activity in ports.

**Validation level 5: consistency with data of other data providers**

Validation level 5 could be defined as plausibility or consistency checks between the data available in the data provider e.g. Government, Other UN Agencies, NGOs working in the same emergency/area.

**Validation rules**

The validation levels, as anticipated in the examples of validation levels, are verified by means of rules. Rules are applied to data, a failure of the rule implies that the corresponding validation level is not attained by the data at hand.

As explained in the previous section, a first broad classification of validation rules distinguishes rules to ensure technical integrity of the data file ( “Structural Validation”) and rules for logical/statistical consistency validation (“Content Validation”). The distinction is useful since the rules used in the two contexts can be very different. Some of them will be presented further below:

**Structural Validation: Rules to ensure technical integrity of a data file format and structure:**

* formal validity of entries (valid data type, field length, characters, numerical range)
* presence of an entry
* no duplicate units
* all the values in a field of one data set are contained in a field of another data set (for instance contained in a code list(s)
* each record has a valid number of related records (in a hierarchical file structure)

Also, rules are often implemented as conditional checks, i.e. they are only checked, if a certain condition holds. This can be regarded as another property of a rule and might be considered as additional “dimension” of the rule typologies (for both rule sets, Categories A and B).

* if “no returnees in the observation unit” (then ‘number of returnees’ should be 0), or
* if “legal form: Self-Employed” (then number of self-employed" must exceed 0), or
* if “no. of employees not zero” (then wages and salaries must be greater than zero), or
* if “enterprise reports production of goods” (then it should also report costs for raw material), etc.

Of course there might be several conditions combined by logical AND or OR statements.

Table below presents at least one example for each rule type in Category A.

**Table: Examples of rules to ensure technical integrity of a data file (Structural Validation)**

|  |  |
| --- | --- |
| Formal Validity of   * Data type * Field Length * Characters * Numerical range | * Telephone Number: Numeric * *Date:* If Date is given as text it should be 8 characters long * Date: If Date is given as text it should contain only numbers. * Month: Month of arrival in the country must be in {1,...,12} |
| Presence of an entry | * Persons in households: It is checked whether all have responded. * Code for Sex: no missing data. |
| No duplicate units | *Site ID*: Each site has a unique ID number, duplicate ID numbers are not allowed within the data set |
| All the values in a field of one data set are contained in a field of another data set (for instance contained in a codelist) “code list check” | Country of origin: Field "country of origin" must contain only entries from a list of valid ISO country codes |
| Each record has a valid number of related records (in a hierarchical file structure)  “Cardinality check” | Number of members of a family: the aggregated number of persons in each family must be equal to the number of individual rows in the data set corresponding to the members of that family |

**Validate outputs**

“This sub-process is where statisticians validate the quality of the outputs produced, in accordance with a general quality framework and with expectations. Validation activities can include:

* checking that the population coverage and response rates are as required;
* comparing the statistics with previous cycles (if applicable);
* checking that the associated metadata and paradata (the process by which the data were collected) are present and in line with expectations
* confronting the statistics against other relevant data (both internal and external);
* investigating inconsistencies in the data;
* validating the data against expectations and domain intelligence”

The checks that are not usually considered as a part of a ‘data validation’ procedure (i.e., the first and the third item

**The data validation process life cycle**

In order to improve the performance of a statistical production process by managing and optimizing the data validation process, it is useful to describe the data validation process life cycle.

**Design phase**

The design of a data validation process is a part of the design of the whole survey process. The data validation process has to be designed and executed in a way that allows for control of the process.

Validation rules should be designed in collaboration with subject matter specialists and should be based on analysis of previous surveys. Consistency and non-redundancy of rules should be verified.

Activity descriptions

* Assess quality requirements for data sets
* Determine satisfactory set of validation rules for the data.
* Assess responsibilities and roles. Document who is doing what; who is responsible for different actions; who is accepting and adopting the validation rules, etc.
* Integrate the data validation process in the overall statistical production process. Design the connections with other phases of the statistical production processes.

**Implementation phase**

Once the data validation process has been designed, it has to be implemented with a parameterization, thoroughly tested, tuned and become productive.

Common definitions and descriptions applied to data validation are required for a common understanding of the whole validation process.

A proper documentation of the validation process is an integral part of the metadata to be published.

Activity descriptions

* Validation rules are formalized and described in a common syntax.
* Determine metrics for data validation rules.
* Testing. Apply validation rules to test data.
* Refinement of validation rules according to the test results.
* Documenting. Data validation rules should be well documented.

**Execution phase**

The result of execution phase is a flag indicating acceptable and not acceptable data, and generally a score measuring the degree of severity of failure.

Activity descriptions

* Data are checked against the validation rules. Validate data against predefined validation rules.
* Summarizing results. It depends on the user of the results (staff, management or methodologist).

**Review phase**

This phase is aimed at continuous improvement of validation process efficacy and data quality.

Improvement of validation rules due to:

* Replacing those that detect few errors by others more powerful
* Replacing those that ‘mislead’: detect errors that are not real errors
* Improvements in validation rules: detecting more possible errors

Activity descriptions

* Analysis of feedback from stakeholders. Feedback gathered in previous phases.
* Analyzing of outcomes from the execution phase. Identified potential problems, errors, discrepancies, detected systematic problems are analyzed in order to decide whether validation rules should be reviewed.
* Identifying and prioritizing problems.

<https://unstats.un.org/unsd/accsub/2008docs-CDQIO/Ses3-Pap3.pdf>

<http://www.ocdqblog.com/home/the-two-characteristics-of-data-accuracy.html>