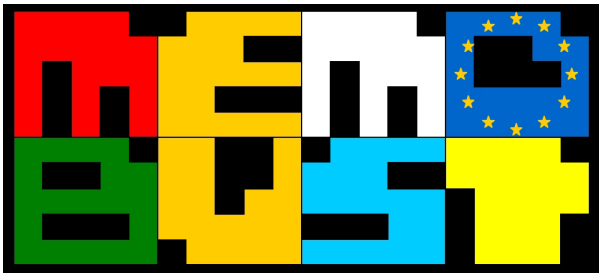

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Logging

Contents

General section

1. Summary

Logging is the activity of producing log information in a log. Log information is used to manage a statistical process and can serve various purposes. These purposes should be determined before implementation of logging. In this theme, we will define logging, logs and log information, and describe various possible areas of application of logging and possible technical solutions for logging.

2. General description

Log information is metadata produced during a specific run of a process. It includes all types of information such as the creation data of an output file, version of the software application that is used and the number of records that is processed. The definition of log information includes quality indicators. So, the definition of log information is quite broad.

Log information can be generated automatically or registered manually. Log information is not metadata that regards the statistical process in general such as process descriptions, methodological documents and descriptive metadata. Statistical output and intermediate results are not considered log information either because these are data and not metadata.

This section discusses the purpose of logging, logging indicators and the quality of logs.

2.1 Purposes of logging

The implementation of logging is dependent on the purpose of logging. Log information is used to manage the quality of the statistical process and output. Logs should be followed up by actions such as validating data (GSBPM sub-processes 5.3 en 6.2), reporting or even improvement of the process, method or system (GSBPM, 2009).

Purposes of logging can be related to the next factors:

1. Punctuality of the output
2. Accuracy of statistical output
3. Traceability and reproducibility of statistical output
4. Statistical confidentiality of statistical output
5. Multiple quality dimensions of statistical output
6. Efficiency of the process

Logs should be designed at the same stage of the design of the statistical methodology and the production system (GSBPM, sub-processes 2.5 and 2.6). Logs are an instrument to manage quality and can be used in validation processes (GSBPM, sub-processes 5.3 and 6.1).

In each subsection, we will describe for each factor how logging can be implemented.

2.1.1 Logging related to punctuality of the output

Some statistical processes have a tight time schedule. If the time to process a dataset takes a few hours or even days it is helpful to know why this process takes so much time. Logging of the performance of

the software that processes statistical data can be used to get insight in possible bottlenecks in the software or database.

2.1.2 Logging related to accuracy of the statistical output

The accuracy of the statistical output is dependent on various factors as mentioned below.

a. Completeness of the units in a micro-dataset

b. Validity of data in a micro-dataset

c. Ability to statistically match units of two datasets

Ad a: The *completeness of units in a micro-dataset* can be measured by logging the number of records in a dataset and compare this number by an expected number. If units are imputed because of incompleteness data about these imputations can be logged too. Several methods for imputation are elaborated in this handbook; see "Imputation – Main Module".

Example: The number of business units in a dataset is 38,000 while 45,000 units were expected.

Ad b: *Validity of data in a micro-dataset* can be checked by applying specific rules (constraints) to the data. This process includes checking on missing values and outliers. Wrong data will be edited. Apart from logging the old value of a variable the violated rule can be recorded. This last information can be used for analysing purposes and as input for improvement of the statistical process. Several methods for editing are elaborated in this handbook; see "Statistical Data Editing – Main Module".

Example: The return of a business unit is Euro 2 million while the number of employees is 400.

Ad c: The *ability to statistically match units of two datasets* can be measured in the process of statistical matching two dataset. Matches and unmatched units can be logged. The ratio between matches and unmatched units is an indicator for the ability to match two datasets. Mismatches (false matches) are, however, hard to discover and cannot be logged. Several methods for matching (or record linkage) are elaborated in this handbook; see "Micro-Fusion – Object Matching (Record Linkage)".

2.1.3 Logging related to traceability and reproducibility of the statistical output

If traceability and reproducibility is required, it is necessary to log the version of the datasets and the version of the software that are used. Moreover, if data are edited or imputed manually it is necessary to log the number of edits and imputations too in order to be able to reproduce the statistical output.

2.1.4 Logging related to statistical confidentiality of statistical output

While analysing the statistical confidentiality, the results of the analysis can be logged. The log could report which details should be or are left out or which data should be or are changed. The log report created for statistical confidentiality is confidential information and should be treated as such. It is only for the NSI concerned, and is accessible by a limited group of persons within the NSI only.

2.1.5 Logging related to multiple dimensions of statistical output

Quality indicators can be regarded as logging indicators. Indicators can cover multiple quality dimension of statistical output and processes. It depends on what indicator is selected.

Example: Quality indicator *item non-response* is an indicator for the accuracy of the output.

2.1.6 Logging related to efficiency of the process

Logs can be used to improve the efficiency of the process.

Example 1: Files can be found more easily as path and filename are logged.

Example 2: Staff capacity needed to run a process manually and automatically can be logged. This log information can be used to analyse if more or less staff should be or can be assigned to the process.

2.2 Logging indicators

There are a number of items that can be logged. It depends on the purpose of the log which items are relevant. Examples of these items are:

- Version of the software
- Version of a file
- Start and completion date and time
- Path and file names of a file
- Time used to create output
- Number of processed records
- Script files. These files make is possible to check if the right procedure is followed.
- Method and rules that has been applied.
- Tuning parameters for a method. Tuning parameters are specified in section 14 of each method module in the handbook.
- Flags: yes/no or more values. A flag indicates for example if a record is edited manually.
- Quality indicators. Quality indicators are specified in section 21 of each method module in the handbook.
- Violated rules: identification of the rule, frequency of violation.

2.3 Quality of logs and log information

Logs and log information should have the right quality too. Quality dimensions of logs and log information are (see the section on the OQRM model in the module "General Observations – Quality and Risk Management Models"):

- Relevance of log information. Log information should be useful and serve a purpose to be relevant.
- Completeness and correctness of log information. If log information is not complete or correct, it could even effect the quality of the statistical output in a negative way.
- Clarity of log information. Log information should be clear to be understood and efficiently used by the user of the log information.

- Accessibility of logs. It should be clear who is authorised to access specific logs.
- Confidentiality of log information. Some logs are confidential such as logs related to statistical confidentiality.

2.4 Example of logging by τ- and μ-ARGUS

τ-ARGUS is a software program designed to protect statistical tables. μ-ARGUS creates safe micro-data files. Both programs produce logs. The "ARGUS Report" contains the following log information (Argus, 2013):

- Creation date of the output table
- Path and filenames of the input files and output file
- Table structure
- Safety rules used
- Time used to protect the table
- Summary of the table, e.g., safe an unsafe cells.
- Version of the software

Separately, τ- and μ-ARGUS both produce a technical log (logbook.txt) that reports the following log information:

- Start date and time of the run
- Version of the software
- Structure of the input table
- Path and filename of the input file
- Start and completion statement

For τ-ARGUS, it is optional to produce a script file that can be used to rerun the program.

3. Design issues

3.1 Beforehand or afterwards

Logging is preferably developed in the design phase (GSPM sub-processes 2.5 and 2.6). However, it could be necessary to produce a log afterwards on an ad hoc basis. A log can be produced, for example, by comparing two files and log the differences. This method can always be used as a last resort if there are no relevant logs available.

3.2 Structure of the log

Logs can be designed as a structured file or as text . In case of a structured file, there are three ways to structure a log file:

- Add a separate file (log) for log information

- Add extra fields to the existing file with statistical data for log information

3.2.1 Separate files

A separate file is useful if the logs concern the dataset as a whole and not separate units in a dataset.

3.2.2 Extra fields

Extra fields are useful if the log information concerns each unit in the dataset. A 'flag' is an example of an extra field. A flag can have two values such as yes or no but is can also contain a code, e.g., a code for a violated rule.

A flag can also be used to generate summery reports. If, for example, a flag is used to indicate that a certain variable is imputed or not, than the total number of imputations can be derived from the flag.

3.3 Presenting log information

Log information can be presented by printing the logs. An alternative is to present the log information on screen and use the theme module in the process of editing for example.

4. Available software tools

Logging is often part of the software that processes the data and seldom a separate software application.

5. Decision tree of methods

A decision tree of methods is not applicable.

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Argus (2013), Website Statistical disclosure control http://neon.vb.cbs.nl/casc/glossary.htm. Retrieved 25 October 2013.

GSBPM (2009), Generic Statistical Business Process Model. Version 4.0 – April 2009. Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS).

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Quality and Risk Management Models
2. Micro-Fusion – Object Matching (Record Linkage)
3. Statistical Data Editing – Main Module
4. Imputation – Main Module

9. Methods explicitly referred to in this module

1. All method modules in the handbook – Section 20: Logging indicators
2. All method modules in the handbook – Section 21: Quality indicators of the output data

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. Phase 2: Design
2. Phase 5: Process
3. Phase 6: Analyse
4. Quality management (as overarching process)

12. Tools explicitly referred to in this module

1. τ- and μ-ARGUS (Argus, 2013). These tools support statistical confidentiality control and produces standard log files. It is used as an example for logging indicators.

13. Process steps explicitly referred to in this module

1. Editing
2. Imputation
3. Statistical disclosure control

Administrative section

14. Module code

General Observations-T-Logging

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	19-03-2013	first draft	Peter van Nederpelt	Statistics Netherlands
0.1.1	11-07-2013	comment SN's reviewers processed	Peter van Nederpelt	Statistics Netherlands
0.1.2	25-10-2013	second round of SN's reviewers processed	Peter van Nederpelt	Statistics Netherlands
0.1.3	10-01-2014	comment HU processed	Peter van Nederpelt	Statistics Netherlands
0.1.4	16-01-2014	log item changed in logging indicator in order to be consistent with the template for methods	Peter van Nederpelt	Statistics Netherlands
0.1.5	03-02-2014	EB's comment processed	Peter van Nederpelt	Statistics Netherlands
0.1.6	04-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:23