Standard Processes

Robbert Renssen, Mattijn Morren, Astrea Camstra and Tjalling Gelsema

Discussion paper (10013)

Statistics Netherlands

The Hague/Heerlen, 2010

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2008–2009	= 2008 to 2009 inclusive
2008/2009	= average of 2008 up to and including 2009
2008/'09	= crop year, financial year, school year etc. beginning in 2008 and ending in 2009
2006/'07–2008/'09	= crop year, financial year, etc. 2006/'07 to 2008/'09 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

6008310013 X-10

Standard Processes

Summary: Statistical processes can be very complex, and it is not uncommon that they are designed and implemented as one big tangle of statistical activities. This paper is an initiative to structure and standardise the processing of statistical data. Concepts like 'standard process step' and 'standard process' are introduced and explained by means of both a non-statistical example (fixing a flat bike tyre) and a statistical example (matching two data files).

Keywords: (statistical) activity, matching, metadata, method rule, process, process step, process rule, product rule, statistical method.

1. Introduction

1.1 Motivation for this document

In the present situation at Statistics Netherlands, lots of local applications are in use, many of which must be considered as a black box. Very few employees are acquainted with these applications and know how they work and what they do. In addition, the maintenance of these applications is very labour-intensive and can only be performed by specialist IT-employees.

Especially due to the expensive maintenance, the management of Statistics Netherlands has formulated the policy to restrict the diversity of tools and applications as much as possible, see Renssen et al. (2008). Preferably, future redesigns of statistical processes should be implemented by the statistical divisions themselves, i.e. without interference of the IT-department, by means of a so-called standard toolset.

However, before standardising the toolset, it is important to get insight into the diversity of the statistical processes and to standardise these processes first. Although every process is unique in some sense, they also have communalities. For example, many statistical processes have in common that they edit data and estimate population parameters. In addition, a common statistical method may be used in editing or estimation.

1.2 Purpose and scope

This paper is about the standardisation of statistical data processes. By means of a number of concepts to be introduced, like *(statistical) activity, standard process step* and *standard process*, the paper suggests a generic structure for processing data. For each particular statistical process, this structure should be further specified. The ideas and concepts are applied to a series of (statistical) activities that results in the record linkage of two data files.

For reasons of understanding, it is important to distinguish between

- Kinds of *(statistical) activities, standard process steps* and *standard processes* taken as generic components.
- Specifications of kinds of *activities, standard process steps* and *standard processes* as applied in specific processes.

In advance of Section 2, an example of a kind of statistical activity is the *selection of units* from a data file. An example of a specification of a kind of statistical activity, i.e. a specific statistical activity, is the *selection of persons older than 15 in the year 2003*.

The purpose of this paper is to structure and standardise the processing of statistical data, in order to make these processes more transparent, flexible, efficient, and easier to document.

- The processes become more transparent because they are designed by means of a generic structure with kinds of *(statistical) activities, standard process steps* and *standard processes* as recognizable generic components (building blocks)
    - If possible, these generic components are based on recognizable statistical methods (standard methodological solution)
    - If possible, these generic components are implemented with standard tools (standard IT-solution)
- The processes become more flexible because the specification of a generic component can be easily adjusted without interference of other generic components.
- The processes become more efficient because the generic components are reusable, and these components are based on standard methodology and standard IT-solutions.
- The processes are easier to document because the generic components need to be documented only once, and henceforth it suffices to document the specification.

The generic structure will be illustrated by means of an elaboration of a standard process (step). This elaboration, however, is not the purpose of this paper. In a series of separate papers we intend to make an inventory of the most important standard process steps[1] and describe these steps according to the lines of the generic structure as proposed in this paper.

[1] A good start of this inventory is provided by the Generic Statistical Business Process Model, see Vale (2009).

1.3 Relation with the Business Architecture of Statistics Netherlands

The Business and Information Architecture of Statistics Netherlands (see Huigen et al 2006) distinguishes between design and implementation of statistical processes on the one hand and between different process stages (data collection, data processing and data dissemination) on the other hand. The generic structure for processing data to be described in this paper can be considered as an elaboration of the Business and Information Architecture in the domain of data processing. Together with the inventory of standard process (steps), it constitutes a framework for designing the processing of data.

1.4 Overview of this paper

Section 2 starts with the introduction of some necessary terminology. The proposed concepts are explained by means of a non-statistical example, namely fixing a flat bike tyre. Using these concept the process of *fixing a bike tyre* is structured. In Section 3 the concepts are applied to a well-known statistical process, namely *matching two data files*. The relation with our series of (standard) statistical methods is discussed in Section 4 and the relation with our standard toolset is given in Section 5. Section 6, finally, gives some general conclusions.

1.5 Related Literature

The basic ideas in this paper are not new. Similar views have been expressed in e.g. Willenborg (2008) and van der Laar (2008). The elaboration of these ideas, however, is new.

In two recent redesign projects the idea of standardisation has already been implemented, see Bruekers et al. (2009) and Reedijk and Kent (2008). Even though these projects use their own interpretation of standardisation, they show that standardisation can work in practice.

2. Processing statistical data

Suppose a statistician is not a statistician, but a bicycle repairer who has to fix a flat tyre. Typical activities include *turning the bike over*, *unscrewing the valve*, *separating the (outer) tyre from the rim*, *removing the (inner) tube*, *sliding the pumped up inner tube through some water (to locate the puncture)*, *marking the puncture, drying and roughing the surface around it, sticking a patch on the hole*, *checking the tyre for sharp objects*, *removing the sharp objects*, *putting back the (inner) tube, putting back the (outer) tyre*, *turning the bike over* and *pumping the tyre*.

For a number of these typical activities the bicycle repairer has a nice piece of equipment. For example, to remove the tyre he/she uses tyre levers and to find the hole he/she uses a bucket of water.

A sequence of linked activities that is triggered by an event and aimed at a particular output by means of converting certain inputs, is called a process. If these activities concern the processing of statistical data and the output is statistical data, then the process is called a statistical production process.

Note that the process of fixing a flat tyre is triggered by a flat tyre (event). The output of this process is a fixed tyre. The description of the repair process only involves the so-called happy flow. If the desired output is not realised according to some stated quality requirements, the process has to be repeated. Alternatively, the process ends and the output is accepted.

Note further that the process of fixing a flat tyre could be part of a larger process, e.g. *repairing a bicycle*, in which the bicycle chain has to be replaced as well. Then this larger process consists of two subprocesses, namely *repairing a puncture* and *replacing a bicycle chain*. Each subprocess has an intended output and consists of a series of related activities to realise these outputs.

Figure 1: a library of typical activities

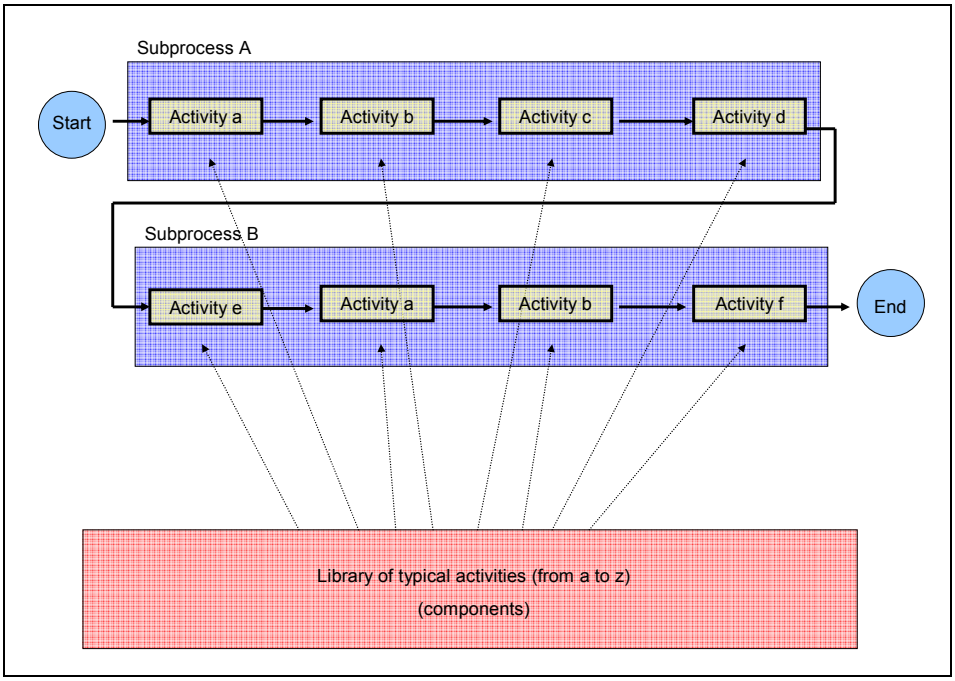


Figure 1 reflects the happy flow of a process. The overall process is subdivided into two subprocesses. Each subprocesses is composed of a number of typical activities, which are taken from a library of typical activities. Activities may be used more than once.

The recognition of typical activities and dividing processes into subprocesses is an important first step in standardising statistical production processes. This idea will be further elaborated in the remaining part of this section.

2.1 Standardisation of production processes

2.1.1 Activities

In this paper, activities are the 'smallest' components we distinguish in modelling and standardising the statistical production process. Like the example of fixing a bicycle tyre, these statistical production activities are more or less inspired in an intuitive way. They describe the production process at a rather abstract level. In principle, these activities could be further decomposed into smaller pieces. However, this often results into a physical description of the production process. Such a description often depends on the equipment that is used to perform the activities. For example, pumping up the tyre can be further physically described as energetically moving up and down the handle of a pump. For the purpose of this paper, it suffices to consider an activity as the smallest component.

An activity is an *autonomous operation* having its own *functionality*. That is, we know what the activity does, irrespectively of its context. For example, the activity *pumping up a tyre* injects an air stream into a tyre, regardless of the reason to inflate a tyre. Examples of typical activities in the domain of statistical processing are

- Selecting units,
- Selecting variables,
- Recoding variables,
- Deriving variables,
- Aggregating over units
- Merging data files
- Sorting data files

Often, these activities are available as pre-programmed modules in statistical software. Especially due to the property of autonomy, activities are very suitable as building blocks for production processes.

2.1.2 Standard process steps

Although using (typical) statistical activities as building blocks for production processes is an important step towards standardisation, it should not end there.

As already said, each statistical activity in a production process has its own functionality, and all these functionalities together contribute to the desired output of the process. Moreover, it is assumed that this output is also the solution for the intended objective of the process.

The desired output of the process fixing a flat tyre is a *fixed tyre*. This fixed tyre is not necessarily the intended objective of the process. This objective probably is *having an undamaged (airtight) tyre*. The solution was to fix the old damaged tyre and this solution resulted in the series of activities that started with *turning the bike over* and ended with *pumping the tyre*. Similarly, the solution could have been to

replace the old damaged tyre by a new tyre. This would have resulted in a somewhat different series of activities and hence into another process. This alternative process has the same objective (having an airtight tyre), yet a different output (a new tyre).

The explicit recognition of the objective of a production process, and specifying and classifying these processes according to these objectives is an important next step in standardisation. In addition, we introduce the concepts of *standard process step* and *standard process* (the latter will be discussed in Section 2.1.5).

We define a standard process step as a sequence of one or more activities. Just like a single activity, a standard process step is *autonomous* and has its own *functionality*. However, in contrast with a single activity, this functionality is always of a statistical nature. That is, the functionality of a standard process step is inspired by a statistical objective and is based on a methodological solution to meet this objective. A standard process step is always controlled by a set of so-called product and method rules (see Section 2.2.1).

- A statistical objective in the processing domain is to enlarge the relevancy, accuracy or coherence of statistical data sets. Statistical objectives are formulated in the design stage. They are expressed by means of the conceptual metadata and quality metadata of a statistical product, see Section 2.3.
- A methodological solution is a systematic solution that is based on one or more statistical methods and/or subject matter knowledge. Methodological solutions are formulated in the design stage. They are expressed by means of (methodological) process metadata, see Section 2.2.3.

It is important to distinguish between a generic solution and a specification of that generic solution. An example of a generic solution to estimate population totals is *weighting using a regression model*. An example of a specification of this generic solution is *weighting using a regression model by age and gender*.

Similarly, it is also important to distinguish between a generic objective and a specification thereof. For example, *obtaining a relevant and accurate micro data set* states a generic objective, see also Section 2.3. An example of a specification of the generic objective is a designed quality version of a statistical product in terms of conceptual metadata and quality metadata. That is, the population delineation is given, as well as the variable definitions, the statistical period and the quality requirements with respect to the accuracy and timeliness (time lag).

Given a generic objective there may be several generic solutions, and hence several generic standard process steps may do the job. Specifying the objective and the solution results in a specific standard process step.

In Section 3 a generic standard process step that is often used in making statistics will be elaborated. This standard process step matches two data files. Before this, the two standard process steps that can be recognized in the process of fixing a flat tyre will be described. These involve *patching the tube* and *inspecting the (outer) tyre*.

2.1.3 Patching the tube

As already said, the objective of the process *fixing a flat tyre* is to have *an undamaged (airtight) tyre*. This objective can be decomposed into *having an airtight inner tube* and *having the (outer) tyre free from sharp objects*. The first objective is met by *patching the tube*. The second objective is met by *inspecting the tyre for sharp objects*. Each combination of objective and solution will be modelled as a standard process step. We start with *patching the tube*.

Of all the activities that are needed to fix a flat tyre, there is one typical activity that is fundamental to obtain an airtight inner tube. This typical activity is *to stick a patch on the puncture.* In order to be able to stick a patch on the tube, the hole should be sufficiently small and the surface around the puncture should be smooth and dry.

Figure 2a: Structure of the standard process step *patching the tube*

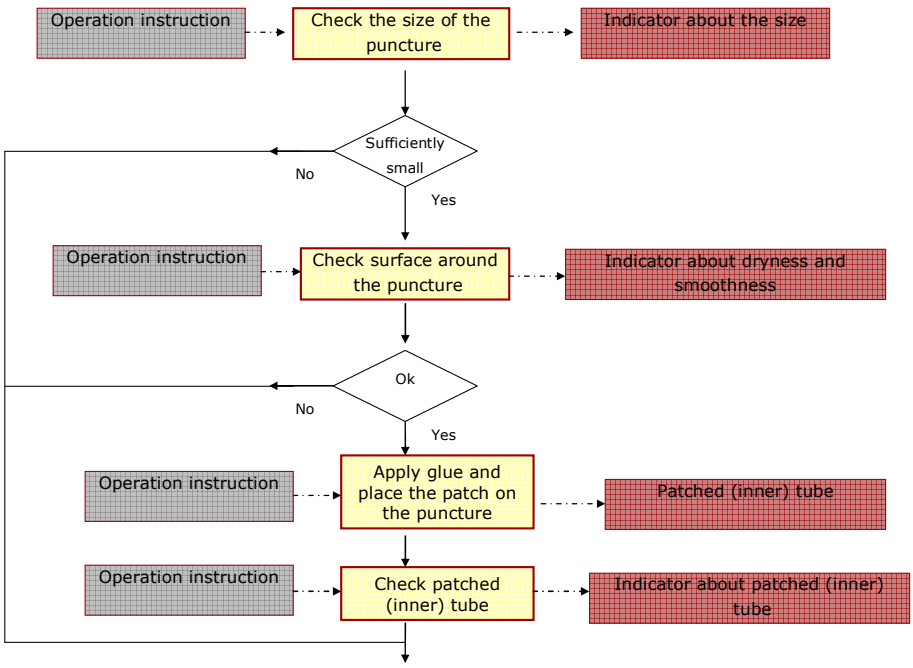
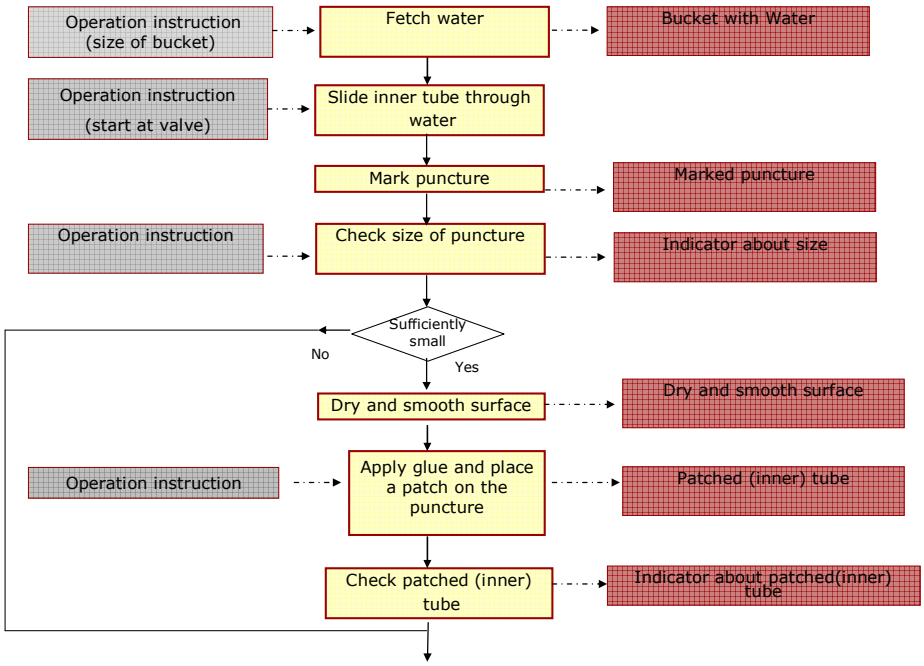


Figure 2a gives a schematic representation of a possible structure of the standard process step *patching the tube*. This structure consists of a minimal number of activities. Naturally, one of these activities concerns the fundamental activity *sticking the patch on the puncture*. Preceding this activity, there are two activities that check pre-conditions necessary to successfully carry out the fundamental activity. If either of these pre-conditions is not satisfied the standard process step ends. Following the fundamental activity there is a activity that checks a post-condition of the fundamental activity, i.e. the quality of the patched (inner) tube. The results of the standard process step are 1) quality measures with respect to the pre-conditions, 2) a patched (inner) tube, and 3) a quality measure with respect to the post-condition.

Alternatively, one could model the standard process step *patching the tube* according to figure 2b. In the alternative model the standard process step also includes the typical activities *fetching a bucket of water*, *sliding the inner tube through the water (to find the puncture)*, *marking the puncture* and *drying and roughing the surface around the puncture.* After all, these preparatory activities are so closely related to mending the puncture that they can be taken together in one step. The results of the extended standard process step still are 1) quality measures with respect to the pre-conditions, 2) a patched (inner) tube, and 3) a quality measure with respect to the post-condition.

Figure 2b: Alternative structure of the standard process step *patching the tube*



There is no clear preference for one of these ways of modelling (figure 2a or 2b), and nothing prevents us from incorporating both structures into our library of standard process steps. The structure of figure 2b may be more complete because it includes the 'methodological solution' of locating the puncture. However, if this solution is not applicable the structure of figure 2a is preferable.

2.1.4 Inspecting the (outer) tyre

Just as there is one bike repair activity that is fundamental to obtaining an airtight inner tube, there is one activity fundamental to having *the (outer) tyre free from sharp objects*. This activity is *removing the sharp object*. In addition, we choose to incorporate the activity *finding the sharp object* in the standard process step *inspecting the outer tyre*, as this is an important pre-condition for removing the sharp object. So, this standard process step consists of two typical activities, the pre-condition *finding the sharp object* and, if found, *removing the sharp object*. The

operation instruction to find the sharp object is to look systematically, starting at the most likely position indicated by the marked puncture.

2.1.5 Standard processes

The two standard process steps *patching the tube* and *inspecting the (outer) tyre* form the heart of the process *fixing a flat tyre*, in the sense that these process steps are essential for realising the objective *having an undamaged (airtight) tyre*.

The question remains whether more standard process steps can be identified in the example of fixing a flat tyre. For example, should the activities needed to *remove the tube* be modelled as a standard process step? Now, if we do so, then the output of this step would be *a removed (inner) tube*. The operation instructions describe how, using tyre levers, to lift the tyre from the rim and to pull the out the tube. The activities that would belong to this standard process step are *unscrewing the valve and lifting it from the valve hole*, *removing the (outer) tyre from the rim*, and *pulling out the (inner) tube*.

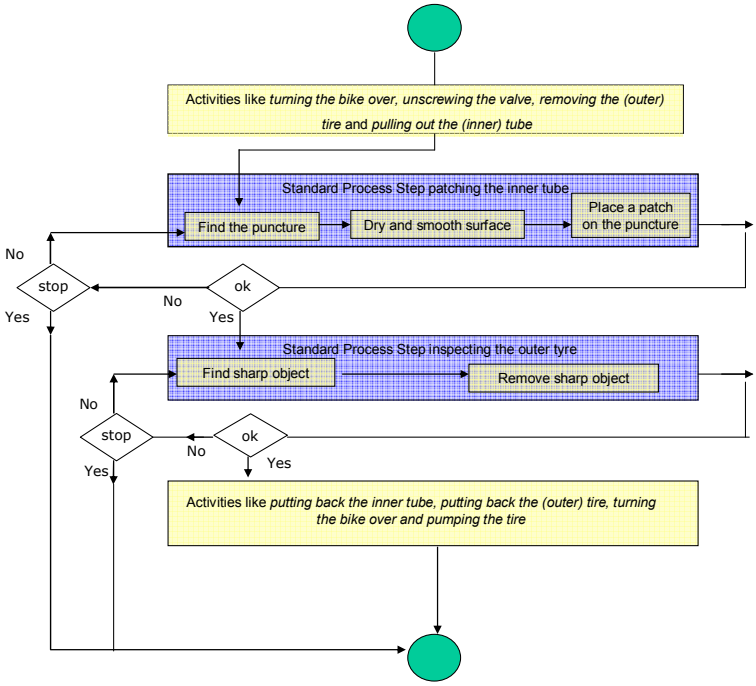
However, the process step *removing the tube* has no added value with respect to the quality of the bike. In other words, this process step has no 'statistical' objective, but fulfils a 'technical' condition that is only necessary for patching the (inner) tube. Therefore we model the activities involved in removing the inner tube as individual preparatory activities and not as a standard process step.

Figure 3 shows the complete process of fixing a flat tyre.

- The process starts with the elementary activities *turning over the bike*, *unscrewing the valve*, *removing the (outer) tyre*, *removing the (inner) tube* and ends with the elementary activities *putting back the inner tube, putting back the (outer) tyre*, *turning the bike over* and *pumping the tyre*.
- The standard process steps *patching the tube* and *inspecting the tyre* are at the heart of the process, and actually realise the objective of the process, that is *having an undamaged (airtight) tyre*.
- When the job is not done well – this could be observed e.g. the next day – the process starts again from the beginning.

We define a standard process as a process that has a specific structure. This structure is dictated by one or more standard process steps whose statistical functionalities achieve a statistical goal. The order and planning of the different standard process steps (and additional activities) are governed by means of so-called process rules see Section 2.2.2.

Figure 3: Example of the standard process *fixing a flat tyre*



2.1.6 The relation between activities, process steps and processes

First, note that the differences between activities, standard process steps and standard processes are gradual. Activities and standard process steps are both *autonomous operations* with their own *functionality* which can be used as building blocks of processes. Contrary to an activity, a standard process step has a statistical functionality that is based on statistical methods and/or subject matter knowledge. The application of these methods and/or subject matter knowledge is formalised by means of product and method rules, see Section 2.2.1.

A standard process is a process that has an explicitly defined *statistical* objective. It consists of standard process steps and additional activities to realise this objective. With reference to Figure 1, standard processes should not only be designed by using a library of activities but preferably by using a library of standard process steps.

2.2 Prescriptive metadata

Like the operation instructions that 'manage' the process *fixing a flat tyre* a statistical production process has to be managed. We distinguish between the management of a standard process step and the management of a standard process. We start with the former.

2.2.1 Management of a standard process step

In performing a standard process step in a statistical production process, two types of guiding rules can be distinguished

- Product rules: This type of rule corresponds to the design of a statistical data set. It describes the metadata of both input- and output data sets at the conceptual level. According to the Business Architecture of Statistics Netherlands, designing a statistical data set consists of specifying 1) the objecttypes and corresponding population delineations 2) the sets of variables per objecttype, 3) the reference period, and 4) the quality requirements.
- Method rules: This type of rule corresponds to the design of the methodological solution at the conceptual level. The Business Architecture of Statistics Netherlands differentiates between three types of method rules
    - Method rules to create the designed output data sets by processing the designed input data sets,
    - Method rules that measure the quality of the output data sets (these are the product quality indicators),
    - Method rules to relate the measured quality to the quality requirements (these are the quality standards).

These types of prescriptive rules result from the design stage of a statistical production process.

2.2.2 Management of a standard process

In implementing a complete standard process two types of process-oriented rules can be distinguished

- Process rules/process design: This type of rule specifies
    - The order in which the specified activities and standard process steps have to be carried out as well as the versions of the product and method rules used,
    - The process (quality) indicators to measure the efficiency and progress of the process,
    - The corresponding quality standards in order to end or (re)iterate (a part of) the process.
- Planning rules: This type of rule indicates the resources and time schedule of the process.

A process rule or design is a time-ordered collection of (versions of) product rules, method rules, and process (quality) indicators with corresponding quality standards.

The process and planning rules also stem from the design stage of a statistical production process. At Statistics Netherlands the design of a process is supported by the standard tool MAVIM.

2.2.3 Process metadata and prescriptive conceptual metadata

The method and process rules together make up the process metadata. The former can be considered as the substantive part of the process metadata (know) and the latter as the process part (flow). By separating the 'know' from the 'flow', the 'know' of the process can be redesigned without altering the 'flow'. In order to design the method rules, methodological knowledge or subject matter knowledge is needed. The product rules constitute the prescriptive conceptual metadata and the prescriptive quality metadata. Once the process is carried out an the output is realised, the prescriptive metadata becomes descriptive metadata.

2.3 Statistical objectives

In Section 2.1 we have associated standard processes with statistical objectives. Many statistical production processes mainly differ in their methodological (and technical) solutions. If we abstract from these solutions, we often see that the underlying statistical objectives are the same. These common objectives offer a nice steppingstone for standardisation.

Irrespective of the chosen solution, many statistical production processes pursue the following 'value chain' of statistical objectives[2].

- Obtaining relevant, coherent and reliable microdata including quality information with respect to the relevancy, coherence and reliability of these microdata[3].
- Obtaining relevant, coherent and reliable estimates for population parameters (aggregated data), including quality information with respect to the relevancy, coherence and reliability of these estimates[3].
- Obtaining relevant, coherent and reliable estimates for population parameters (aggregated data) that are consistent with conceptual restrictions from so-called integration systems. The conceptual description of these integration systems includes quality information regarding the relevancy, coherency and reliability of these estimates[3].
- Obtaining relevant, coherent and reliable time series, again including quality information with respect to the relevancy, coherence and reliability of these estimates[3].

The objectives mentioned above can be considered as final objectives that are of interest to external users of statistics. These objectives follow from the mission statement of Statistics Netherlands that says 'to publish reliable and coherent statistical information that anticipates the statistical needs of society'. The statistical

---

[2] Note that these types of statistical objectives are formulated in terms of statistical products. They are not formulated in terms of statistical activities.

[3] This could be information about the processing of the data, if this kind of information is important to interpret the quality.

objectives are concretely specified in the design stage of a statistical production process resulting in a conceptual product design, see Section 2.2.

Each product design eventually fulfils an external need. However, since many statistical processes are organised as a chain (or network) of local processes, the output of a local process is not necessarily a final product. In case of intermediate products there should be an internal need for the statistical information. Intermediate products differ from final products because

- The population delineations and/or the definitions of the variables still need adjustment, i.e. the relevancy and coherence standards are not yet met,
- The consistency and/or accuracy of the data still need improvement, i.e. the reliability standards are not yet met,
- Statistical disclosure is not yet sufficiently assured.

Depending on the situation (available budget, available data sets, time schedule, etc.), each (local) statistical production process follows its own methodological solution to realise the product designs.

As an example, the methodological solution *macro editing* to obtain sufficiently reliable microdata corresponds to another production process, i.e. series of activities, than the methodological solution *automatic micro editing*. Another example concerns the measurement of overcoverage of a microdata set as a quality indicator. Overcoverage could be measured by matching the microdata to a population frame and subsequently counting the number of failed identifications. Alternatively, it could be measured by counting the number of failed domain controls.

The difference between an objective of a statistical production process and the output of a statistical production process can be expressed as follows

- A statistical objective refers to the product design in terms of prescriptive conceptual metadata and prescriptive quality metadata. There is no methodological solution involved.
- Statistical output concerns the realisation of the product design and is characterized by descriptive conceptual metadata and descriptive quality metadata. The realisation is based on a methodological solution and is governed by product rules, method rules and process rules.

The difference between design and realisation can be considered as a measure for the effectiveness of the methodological solution, and hence of the process.

3. Matching as a standard process (step)

When different statisticians are asked to give their understanding of the procedure *matching data sets* they would probably give different answers. Some answers would be close to our notion of *standard process step*, while others would more

closely resemble our notion of *standard process*. In this section both views are modelled. First, we give some background information on matching.

3.1 Background and definition

The functionality of matching two data sets is often used for the purpose of obtaining relevant, coherent and reliable microdata, including quality information [4].

- Matching two data sets contributes to relevancy and coherence by putting together two separate sets of variables and thereby providing additional information about the relationship between these two sets of variables.
- Matching two data sets contributes to the reliability by measuring and correcting for coverage errors.

Consider two micro data sets with a common population and a number of common variables and further assume that one of these data set is the recipient data set and the other data set is the donor data set, the objective stated above is realised as follows:

- Based on the common variables and a matching criterion, the units in the donor data set are matched with the units in the recipient data set.
- The number of positive matches is measured (i.e. the unit belongs to both the recipient and the donor data set).
- The number of missed matches is measured (i.e. the unit belongs to the recipient data set, not to the donor data set).
- The number of missing matches is measured (i.e. the unit belongs to the donor data set, not to the recipient data set).
- The variables to be added are selected and associated with the recipient units.
- Depending on the quality of the donor data set and the matching criterion the missed donor units could be added to adjust for undercoverage, and the missing recipient units could be deleted to adjust for overcoverage.

The common variables that are used to identify and relate the recipient and donor units are called matching variables. Ideally, these matching variables have identical definitions and provide unique keys for identification. Unfortunately, these matching variables often have similar but not quite identical definitions. For example, there may be slight differences between matching variables due to time lags or different ways of rounding numbers. In addition, the matching variables may contain slight errors or even missing data.

In case of imperfect keys, one needs a so-called matching criterion to decide whether two keys are ‘close’ enough for a positive match.

---

[4] We only discuss exact matching. We consider statistical matching (or synthetic matching) as a different type of statistical activity.

Suppose that two micro data sets referring to a common population of persons, should be matched using five matching variables. The recipient data set contains the variables Social-Security-Number, Name, Place-of-residence, Gender and Age. The donor data set contains the variables Social-Security-Number, Name, Place-of-residence, Gender and Date-of-birth. Suppose further that the variables Social-Security-Number and gender are identically defined (but may contain missing values and errors). Finally suppose the variables name and place-of-residence are open text variables (these variables may also contain missing values and errors).

Provided that the variable Social-Security-Number is the perfect key in both data sets, i.e. there are no missing values and/or errors, this variable is ideally used to match the recipient and donor units. Therefore, the validity of the social security numbers is checked. For valid social security numbers the recipient and donor units are actually matched, using as a matching criterion that the numbers be identical.

Suppose the number of positive matches is insufficient because of too many missing values in the variable Social-Security-Number. The unmatched recipient and donor units may be matched in a second iteration, using the matching variables Name, Place-of-Residence, Gender and Age/Date-of-Birth. Now, in order to use the open text variables name and Place-of-Residence, these variables have to be coded first. That is, by means of a so-called knowledge table these open text variables have to be transformed into new variables having a finite number of categories. In order to use Age and Date-of-Birth as matching variables, the variable Age in the donor data set has to be derived from the variable Data-of-Birth. Finally, it could be necessary to recode the categories of the variable Gender, e.g. from {male, female} into {0,1}.

Provided these (re)coded and derived variables have no missing values and/or errors – and to that end they are checked on validity – these variables can be used as matching variables for the second iteration. An example of a matching criterion would be that the values of the variables Name and Place-of-Residence should be identical in both datasets and also that there is a positive match on either Age or Gender.

Given a recipient unit, there are three possibilities: 1) there is no matching donor unit found, 2) there is exactly one matching donor unit found, and 3) there is more than one matching donor unit found. The first possibility is called a missed match, the second and third possibility a positive match.

The probability of several matching donors and hence of assigning the wrong donor unit to a recipient unit can be judged by a frequency distribution with the set of matching variables as classification variables.

When matching two data sets we obviously need two data sets as input, namely the recipient data set and the donor data set. Furthermore we need a matching criterion. The logical data models of the data sets may be complex, i.e. both data sets may refer to more than one objecttype, as long as they have

- a common population delineation (and hence a common objecttype), and
- a common set of variables with respect to this objecttype.

Figure 4a and 4b illustrate a recipient and donor data set that satisfy these conditions. The resulting matched data set is illustrated in figure 4c.

Figure 4a: Example of a logical data model of the recipient data set

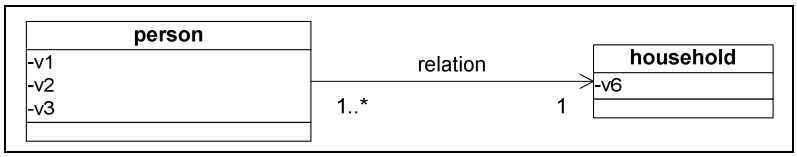


Figure 4b: Example of a logical data model of the donor data set

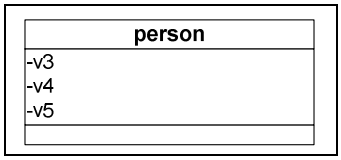
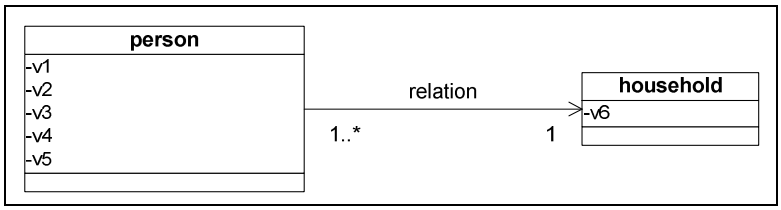


Figure 4c Example of a logical data model of the matching result



The population delineation of this matched data set is determined by the common population of the input data sets. The set of variables with respect to this common population is a selection of variables from both the recipient and donor data set. Note that the specific population delineation can be taken as

- The intersection between the recipient and donor data set (inner join). This population is operationalized by the positive matches.
- The recipient data set (left join). This population is operationalized by the positive matches and the missed matches.
- The donor data set (right join). This population is operationalized by the positive matches and the missing matches.
- The union between the recipient and donor data set (outer join). This population is operationalized by the positive matches, the missed matches and the missing matches.

The choice for one of these population delineations is a design decision depending on the quality requirement of the output, which is laid down in the product rule.

3.2 Types of activities when matching two data sets

Consider matching two data sets by means of a set of common variables. Starting from the population delineation of the recipient data set, the following activities may be distinguished: reading the input data, *selecting the recipient and donor data set*,

(re)coding variables, deriving variables, *selecting a set of matching variables*, *checking the value range of the matching variables*, *identifying and matching the donor units*, *selecting the variables to be matched for the output*, writing the output data to a file.

3.2.1 Defining the standard process step

The activities listed above in italic can be considered as typical matching activities. These activities are at least part of the standard process step *matching*. In order to monitor the quality of the match, the standard process step also includes a type of aggregation activity to indicate the number of positive, missed and/or missing matches.

Figure 5: Typical structure of the standard process step *matching*

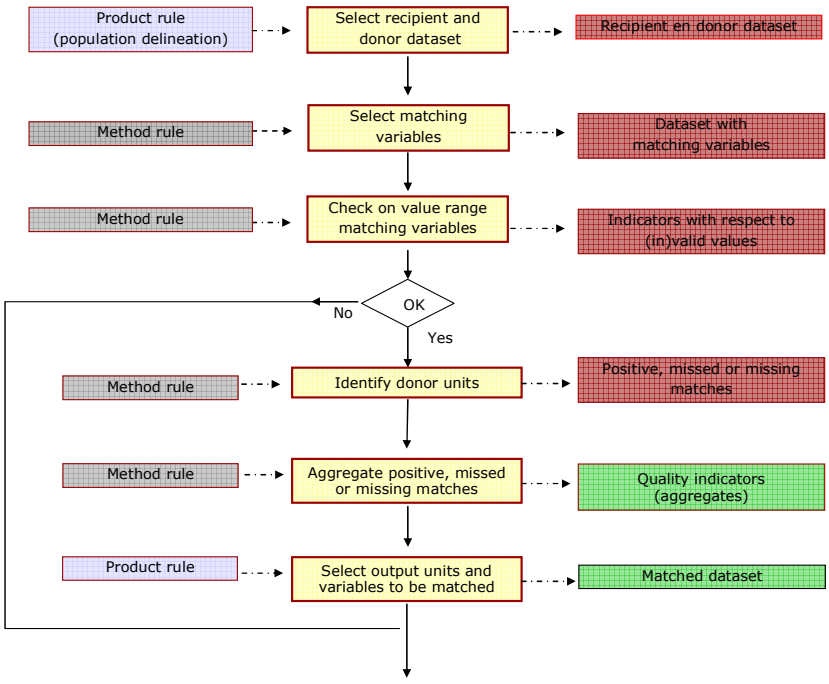


Figure 5 exhibits the generic structure of the standard process step *matching*. The design of this standard process step consists of the following parts

- Design of the product rules
    - Design of the output data set (in terms of population delineations, sets of variables, and quality requirements with respect to over- and undercoverage), see also figure 4c.
    - Design of the recipient and donor data set (in terms of population delineations, sets of variables, and quality requirements with respect to over- and undercoverage), see also figure 4a and 4b.

- Design of the method rules
    - Design of the matching variables (including their range of valid values) and the corresponding matching criterion
    - Design of the quality indicators to measure over- and undercoverage (based on the number of positive, missed and/or missing matches).
    - Design of the standards for the quality indicators in relation to the quality requirements with respect to over- and undercoverage.

The matching criterion may be based on a statistical method. Such a method will be discussed in Section 4.

The implementation of the standard process step *matching* consists of six activities:

- Selecting the recipient and donor data set that is prescribed by the product rule,
- Selecting the matching variables that are prescribed by the method rule,
- Checking the value range of these matching variables,
- Identifying the donor units that match the recipient units,
- Aggregating the positive, missed and/or missing indicates matches to measure the success rate,
- Selecting the output data set that is prescribed by the product rule, i.e. selecting the positive, missed and/or missing matches, as well as the set of the variables to be matched.

Often, the concept *matching* is interpreted as the total collection of typical activities as described in figure 5.

3.2.2 Defining the standard process

The standard process step *matching* as described in the previous section has two important preconditions that should be examined at the design stage. That is, the recipient and donor data sets should have identical variable definitions for the matching variables and they should refer to the same population delineation.

Most recipient and donor data set probably do not exactly satisfy these preconditions and have to be 'prepared' before any matching activities can take place. For example, in order to obtain identically defined matching variables it may be necessary to adjust either the variables in the donor data set or the variables in the recipient data set. We distinguish between

- Coding variables. This involves complex operations in which open text variables are transformed to closed categories (as prescribed by a method rules),
- Deriving variables. New variables are created from other variables by means of deterministic operations based on conceptual variable definitions (and prescribed by a product rule),

– Recoding variables. This involves simple deterministic operations changing the codes of the value range of a variable (the conceptual variable definitions including the conceptual value ranges, remain unchanged).

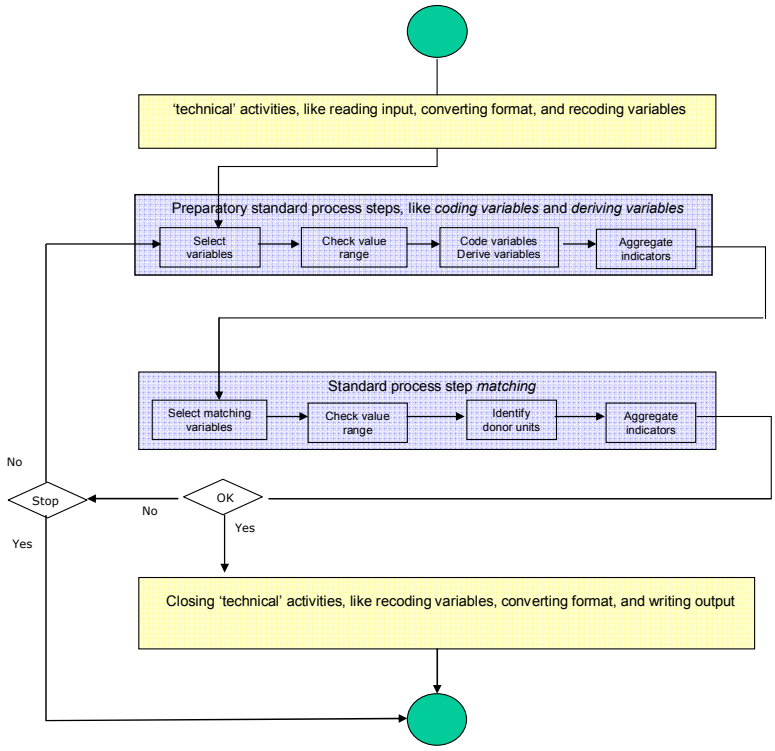
Note that the activity *recoding variables* has no statistical functionality. There is no statistical value added to the data set by this activity. No substantive knowledge is needed to design this activity nor a statistical method. Often, the activity is implemented to fulfil a 'technical' (non-statistical) condition. It merely changes the physical representation of a value range and hence of the data set.

In this paper the activities *coding variables* and *deriving variables* are regarded as standard process steps. These types of activities alter the statistical value of a data set, and one needs a statistical method or subject matter knowledge to design these activities.

Besides the 'preparatory' activities mentioned above, general 'technical' activities are needed, such as *reading a data set from a file*, *transforming the format of a data set* and *writing a data set to a file* to complete the matching process. Figure 6 shows a typical structure of the standard process *matching*.

This standard process consists of a number of standard process steps and a number of supporting 'technical' activities. The essential part of this standard process is given by the standard process step *matching*. As an illustration, the figure also includes a stopping criterion and an iteration criterion.

Figure 6: Typical structure of the standard process *matching*



For each (re)iteration the standard process steps need to be specified, and specifications may differ between iterations. The stopping criterion may be quality driven, time driven, or based on a maximum number of iterations.

Alternatively, the concept *matching* is not only interpreted as the standard process step but as the complete standard process as described in figure 6.

4. A statistical method for the matching criterion

In Section 3.1 two matching criteria were discussed in the framed example. The first matching criterion was a criterion to compare Social Security Numbers. According to this criterion, there is a positive match only if the Social Security Numbers of donor and recipient are identical. The second matching criterion concerned a criterion to compare the combination of the variables Name, Place-of-Residence, Age and Gender in the donor and recipient data sets. This criterion indicates only a positive match if the values of both Name and Place-of-Residence are identical in the two datasets and in addition the values of either Age or Gender.

Now, given a set of matching variables $\{v_1,...,v_K\}$, both criteria can be written as a parameter of a general matching formula

- Define for each matching variable v_k a distance measure D_k to measure the difference between the recipient and donor value,
    - For nominal matching variables D_k could be taken 0 if the values are identical; otherwise D_k is 1.
    - For interval or ratio variables D_k could be taken 0 if the values differ less than $x\%$; otherwise D_k is 1.
- Define for the complete set of matching variables a weighted distance measure $D=\sum_k w_k D_k$ with $\sum_k w_k = 1$. The weight w_k represents the contribution of the matching variable D_k to the matching criterion.
- A match is called positive only if $D \le k$.

Obviously, the first matching criterion can be considered as a special case of this general criterion by defining only one matching variables, i.e. $K=1$ and by taking the weight of this variable equal to 1 and $k=0$.

Also the second criterion can be considered as a special case. Note that $K=4$ and assign the weight $⅓$ to the variables Name and Place-of-Residence, and the weight $⅙$ to the variables Gender and Age. Then, by taking $k=⅙$, the second criterion is obtained.

As already said, the matching criterion has to be defined at the design stage resulting in a method rule. If this method rule can be specified by choosing specific values of

the parameters K, w_k, x and k, then this method rule is called parametric (adaptable by changing the parameters). So, a parametric method rule represents a more general statistical method that can be reused by adapting the parameters.

5. Relation with standard tools

Until now we have discussed the standardisation of processes at the conceptual level. We have made a distinction between generic descriptive rules, standard process steps and standard processes on the one hand and specifications of these generic prescriptive rules, standard process steps and standard processes on the other hand.

Once the standard process steps and processes are specified (by specifying the relevant rules at the design stage), they have to be implemented, and we arrive at the physical level. It is not the purpose of this paper to discuss and compare tools for particular processes such as matching. However, the paper may help in making a choice of a suitable tool.

For example, when selecting a tool to match two data files, the following two questions are important

- What is the scope or functionality of the matching process (standard process step or standard process)?.
- How adjustable is the method rule (parametrised or not)?

Depending on the answers to these questions, three classes of (standard) tools can be distinguished:

- A (standard) tool may be generic with respect to its functionality, but limited with respect to the adaptability of the method rule. Often, such a tool contains a library of types of activities by means of which specified (standard) processes can be built.
- A (standard) tool is generic with respect to the adjustability of the method rule, but limited with respect to its functionality. Often, such a tool supports a certain type of a standard process step, which can be specified by adapting parameters through an interface of this tool.
- A (standard) tool is generic with respect to both the adjustability of the method rule and its functionality.

For example, many general (standard) tools provide a functionality that can be used to match two data files by means of a unique matching variable. In such situations, the adjustability of the method rule is limited to specifying the matching variable. For more complex situations, i.e. situations that require an adaptable matching criterion or situations that need some coding first, such general (standard) tools may fail. In that case, (standard) tools having a generic matching functionality are more qualified.

6. Summary and future developments

In practice, statistical production processes are often designed and implemented in a relatively unstandardized manner as a large collection of activities. This paper suggests a way to standardise these processes by recognizing and defining autonomous building blocks with typical statistical functionalities. These functionalities are based on statistical methodology or subject matter knowledge. The building blocks are called standard process steps.

In addition, the paper suggests that these building blocks are used as much as possible in designing so-called standard processes. A standard process is a process that has a specific structure. This structure is dictated by one or more standard process steps, whose statistical functionalities achieve a statistical objective.

The idea is to compose a library of standard process steps and standard processes as building blocks, and to use this library to design production processes. Then, designing a production process comes down to specifying the building blocks that are needed.

Constructing such a library includes 1) recognizing suitable building blocks and 2) elaborating these building blocks. As an example, this paper discussed the building block of matching two data files.

The paper also suggested a way to recognize building blocks. Building blocks should be based on general methodological solutions that are inspired by typical statistical problems to realise typical statistical objectives.

Future research will be aimed at elaborating other building blocks such as editing and imputing data, estimating population totals, coding open text variables, micro-integration, and so on.

If different building blocks are implemented by different (standard) tools, then there might be a so-called integration problem, in the sense that different tools might not be able to ‘intercommunicate’. In that case additional (technical) activities may be needed to overcome this problem. A next step to standardise production processes would be to recognize common patterns of building blocks, and define these patterns as fixed combinations in which the integration activities are incorporated. Naturally, these fixed combinations can be considered as new building blocks.

Incidentally, the integration problem can be reduced to a large extent if the tools concerned communicate by means of a common metadata model. The integration problem is thus limited to the connection of each (standard) tool to this metadata model.

References

Bruekers, N., Leerintveld, B., van Bracht, E. en de Jonge, E. (2009), Joules in Beweging (JiB); Software Architectuur Document, Versie 1.1p3 (in Dutch).

Huigen, R., Bredero, R., Dekker, W., and Renssen, R. (2006), Statistics Netherlands Architecture; Business en Information model, Internal CBS-report, Statistics Netherlands.

Reedijk, J. en Kent, J. (2008), Finish It; Software Architecture Document, versie 1.0 (in Dutch)

Renssen, R. (2009), DSC voor CBS-brede uitwisseling van statistische data, Centraal Bureau voor de Statistiek, Heerlen (in Dutch)

Renssen, R., Wings, H. en Paulussen, R. (2008a), *Processes, methods and tools*, Centraal Bureau voor de Statistiek, Heerlen.

Renssen, R., Wings, H. en Paulussen, R. (2008b), *Het vervolg van processen, methoden en tools*, Centraal Bureau voor de Statistiek, Heerlen (in Dutch).

Van de Laar, R. (2008), Conceptuele Typering van Processtappen naar Business Functie, Interne CBS-nota, 7 juli 2008, versie 1 (in Dutch)

Willenborg, L. (2008), Pre- en postcondities bij standaard processtappen, Interne CBS-nota (concept; in Dutch).

Windmeijer, D., Projectgroep Architectuur (2006), ICT-Masterplan, CBS Architectuur, Logische Informatie Architectuur (BI002); Eisen mbt Toekomstige BI-Architectuur, Interne CBS-nota, Centraal Bureau voor de Statistiek (in Dutch).