

Process model for editing

Pauli Ollila

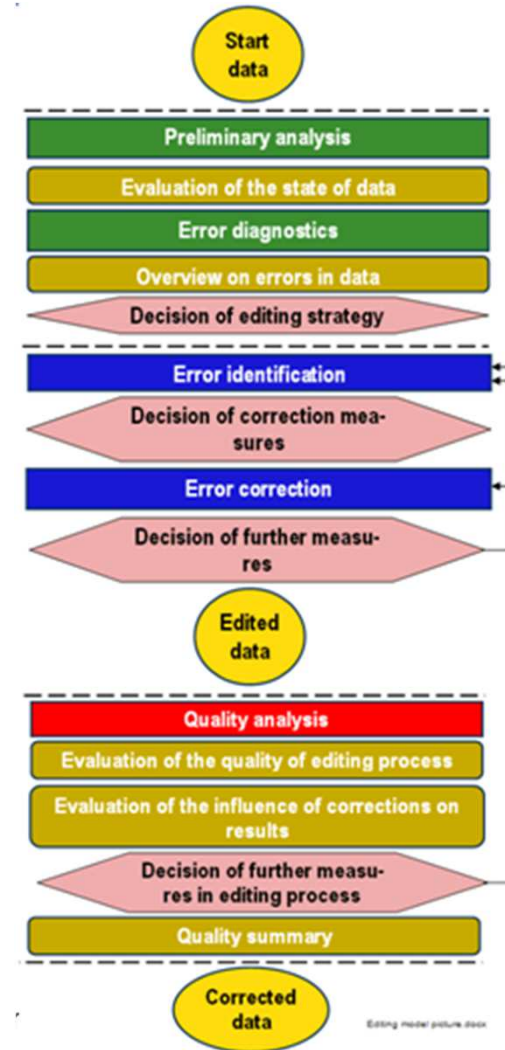
Statistics Finland

Contents:

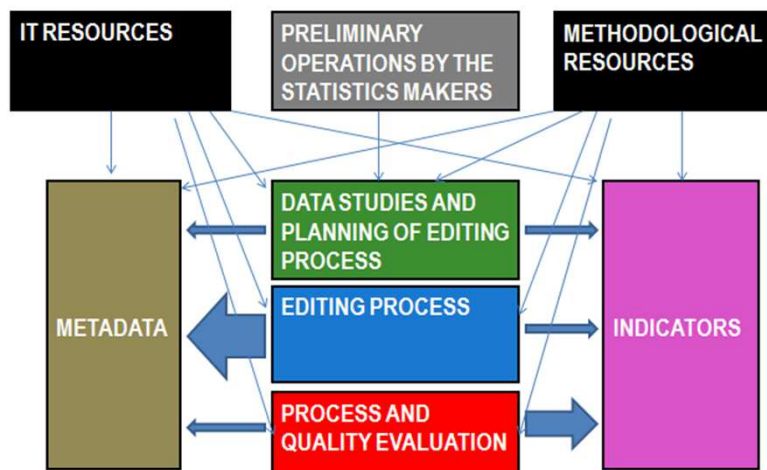
1. Challenging world of making statistics



2. Process model for editing



3. Realisation of the process model: methodologies, practices, IT solutions



1. Challenging world of making statistics



**Errors,
missingness and
uncertainties of
statistical data**

Random errors

Improper responding

Data available in form
not suitable for
response

Don't know in which variable the
error is

Systematic errors

Misunderstanding

Carelessness in
responding

Holes in data
matrix

Too challenging tasks
for respondents

Linking problems of
data sources (variable
definitions,
contradictory values,
varying quality
requirements)

Response describing
wrong theme or unit

Respondent's
definitions or
classifications
problematic

Is this error or not?

Requirements for data and results

- **Consistency requirements** for data and results
 - **Constraints:** some variable entities must fulfill conditions (e.g. partial sums to total sum in Structural Business Statistics)
 - **More general consistency expectations:** variable entities must be in a "sensible range" → danger of subjective and varying corrections without clear reasoning
- **Fullness requirements** for data
 - In some data no item nonresponse can exist (e.g. further use of data)
 - It may happen that there are no clear rules for when item nonresponse is allowed in the final data → varying correction practices
- **Quality requirements**
 - Sometimes **quality requirements for results** are set, and they must be taken into account in data handling during processing
 - **Quality of editing and imputation** is rarely studied, and almost always there are no criteria set for that
- **Time table requirements**



RISKS AND CHALLENGES IN REALISATION

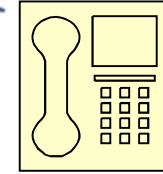


CHOICES OF METHODS

- Not enough knowledge on what editing methods could be carried out in what data / error situation
- Afraid of choosing "difficult" statistical method because of lacking knowledge and lack of existing methodological support
- The simplest imputation alternative is often chosen and there is no background work for improving the quality of imputation (e.g. nonresponse analysis)
- No theoretical information available for testing different editing methods and how to carry out them with software

DECISIONS ON CONDUCTING EDITING

- Timetables and work resources create limits
- Historic data and reference data, how easily they can be utilized
- In which phases which actions should be carried out: data receiving phase, separate data, combined data
- In some tasks there are no tailor-made IT tools, in some tasks tools can be available, but there is no experience of use → choosing a simple method (not necessary bad as such)
- Lack of information of good E&I practices
- Uncertainty of the sensible order of carrying out E&I operations



CARRYING OUT ERROR RECOGNITION

- **Too much manual error recognition** (due to error lists) and lack of **evaluation of cases**: "drowning into" the sea of observations and variable values to be checked
- With an entity of many variables the **error localization** may cause trouble: "lost in error situation", no certainty where the error is
- The routine of decisions lacking in different error situations → choosing the same action as always before

CARRYING OUT FURTHER ACTIONS AND ERROR CORRECTION

- Much too often fetching values / making callbacks also in nonsignificant cases
- In some statistics laborious comparison and data fetching processes
- Lacking systematic style of further actions, needless returns in editing and imputation
- Corrections dispersed in different programs or applications, corrections written every time as program lines
- Problems with coordination of corrections / imputations and realizing constraints
- During operations there is no statistical information about corrections and their effect on results

2. Process model for editing

At first some background ...

From **Steven Vale's** presentation in UNECE meeting 2011, Ljubljana, Slovenia

Generic Statistical Business Process Model

Process



Phases

Quality Management / Metadata Management

Sub-processes

(Descriptions)



1 Specify Needs	2 Design	3 Build	4 Collect	5 Process	6 Analyse	7 Disseminate	8 Archive	9 Evaluate
1.1 Determine needs for information	2.1 Design outputs	3.1 Build data collection instrument	4.1 Select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Define archive rules	9.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Manage archive repository	9.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design data collection methodology	3.3 Configure workflows	4.3 Run collection	5.3 Review, Validate & edit	6.3 Scrutinize & explain	7.3 Manage release of dissemination products	8.3 Preserve data and associated metadata	9.3 Agree action plan
1.4 Identify concepts	2.4 Design frame & sample methodology	3.4 Test production system	4.4 Finalize collection	5.4 Impute	6.4 Apply disclosure control	7.4 Promote dissemination products	8.4 Dispose of data & associated metadata	
1.5 Check data availability	2.5 Design statistical processing methodology	3.5 Test statistical business process		5.5 Derive new variables & statistical units	6.5 Finalize outputs	7.5 Manage user support		
1.5 Prepare business case	2.6 Design production systems & workflow	3.6 Finalize production system		5.6 Calculate weights				
				5.7 Calculate aggregates				
				5.8 Finalize data files				

Why do we need a model?

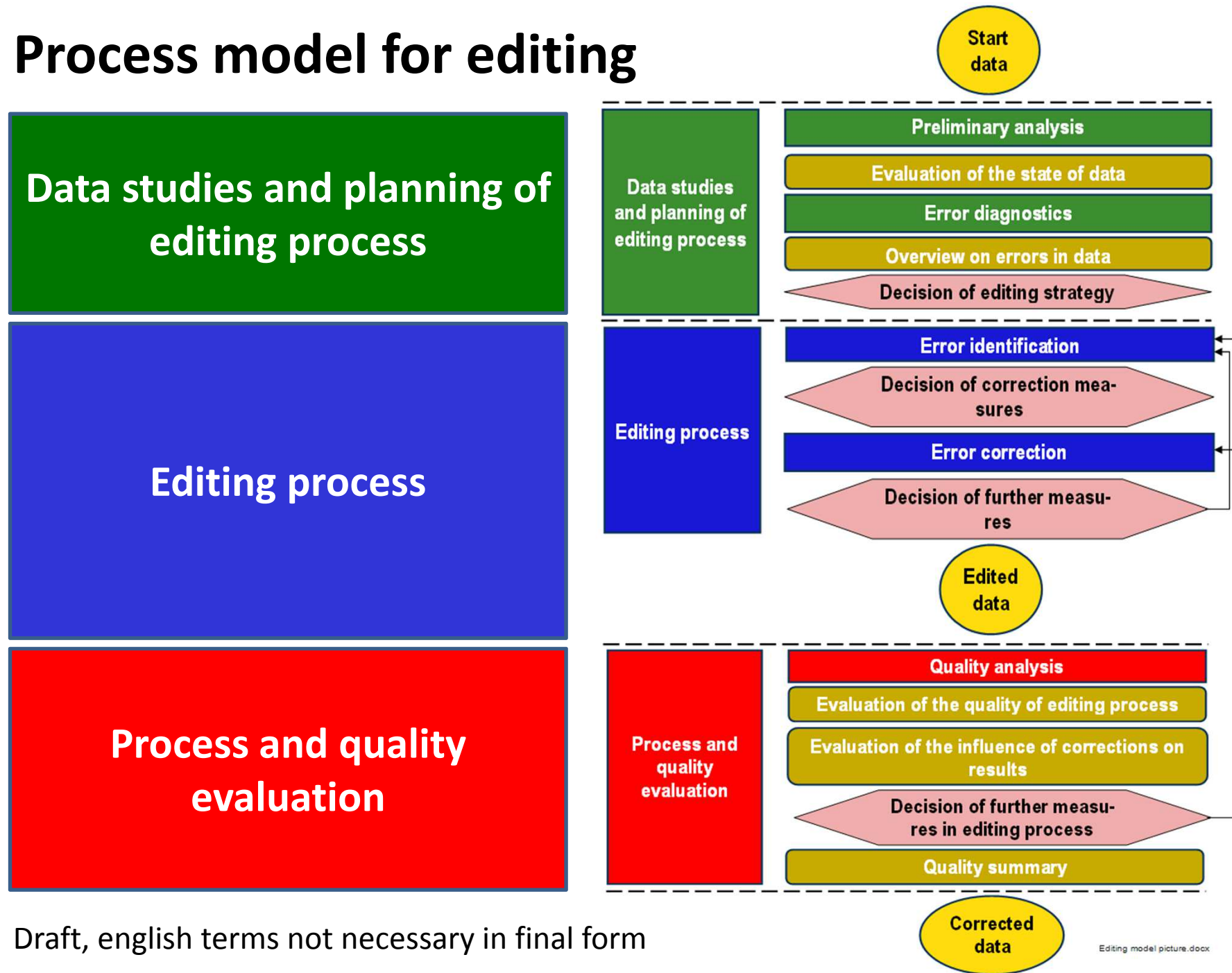
- To define and describe statistical processes in a coherent way
- To standardize process terminology
- To compare / benchmark processes within and between organisations
- To identify synergies between processes
- To inform decisions on systems architectures and organisation of resources

From **Li-Chun Zhang's** presentation about **industrialization of editing** in UNECE meeting 2011, Ljubljana, Slovenia

Standardization as key approach: Getting rid of all *unnecessary* variations

- Production processes: GSBPM
- Production/information standards: GSIM to-be
- Methodology:
 - Standardization ≠ Single solution
 - Recommended/common solution = point-of-departure
 - Deviation allowed, if justified = impetus for auditing
 - Over time, standardization generates a method library
- IT:
 - General statistical systems
 - Architectural design
 - Engineering principles

Process model for editing

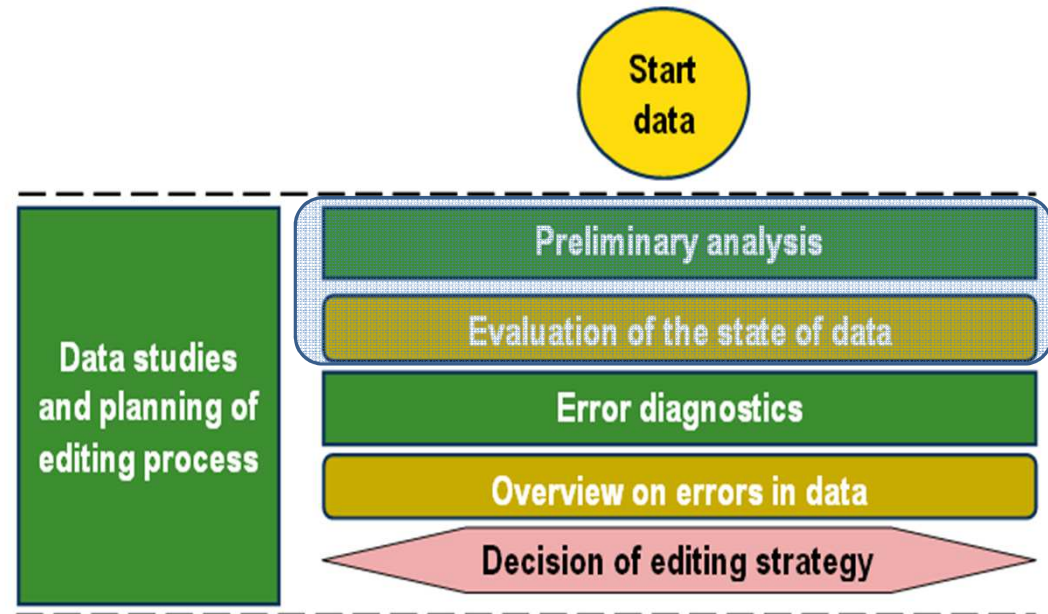


Draft, english terms not necessary in final form

Data studies and planning of editing process

PRELIMINARY ANALYSIS

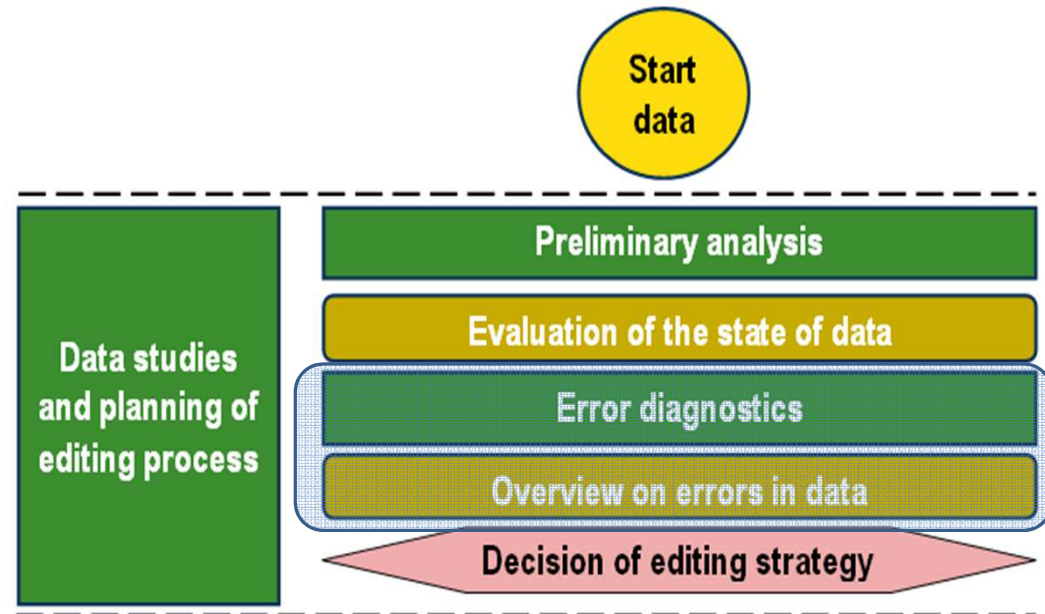
- Gives an overview on the substance state of current data (raw or partially processed)



- **Analysis based on prepared programs** is tabulation and calculation of statistics with relevant subgroups targeted to variables essential for editing process. Some statistics can be defined as "State of data" indicators, which can be calculated at subsequent phases as well for evaluating the development of editing (resembling Canada's "rolling estimates").
- **Interactive data study** is interactive analysis based on the experiences of the researcher using suitable IT solutions (analysis methods, graphical methods, observation value views) → might catch those (possibly new) characteristics, which cannot be found with prepared programs or further studies are needed based on prepared program studies

Data studies and planning of editing process ERROR DIAGNOSTICS

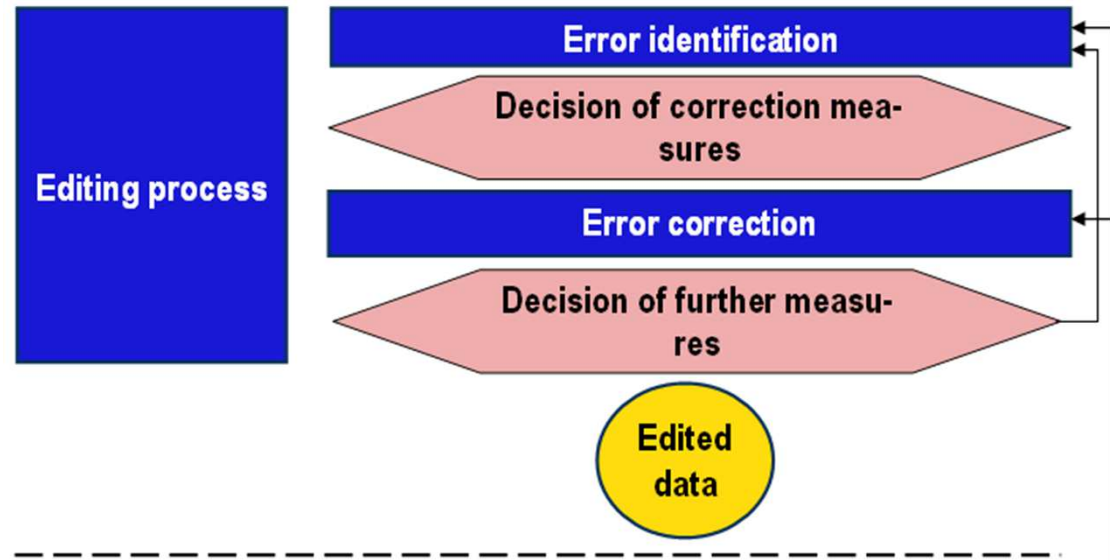
- Making an overview on typical errors in the data and possible changes in the error profile of the data.
- For this phase there should be patterns of useful study practices in different data contexts.



- Here the error identification (this variable value in this particular observation is erroneous) is not the goal, though in some cases the errors could be identified.
- **Analysis based on prepared programs** includes tabulations of fatal errors and clear suspicions found in the data.
- **Interactive data study** is (as in preliminary analysis phase) interactive analysis based on the experiences of the researcher using suitable IT solutions (analysis methods, graphical methods, observation value views). At this phase the goal is to find errors (e.g. systematic), which could not be revealed with previous error procedures.
- Without this phase the new development of error recognition occurs only when new kind of errors are noticed in the editing phase, quite often by chance.

Editing process (in general)

- All editing is realised in the phase of editing process



- Editing process can include several error identification and correction actions → **iterative**
- *Error identification* includes actions, which result to identifying certain and possible error at the observation level or at the group of observations level.
- *Error correction* realises corrections of all or some identified errors following the decisions made at the error identification phase.

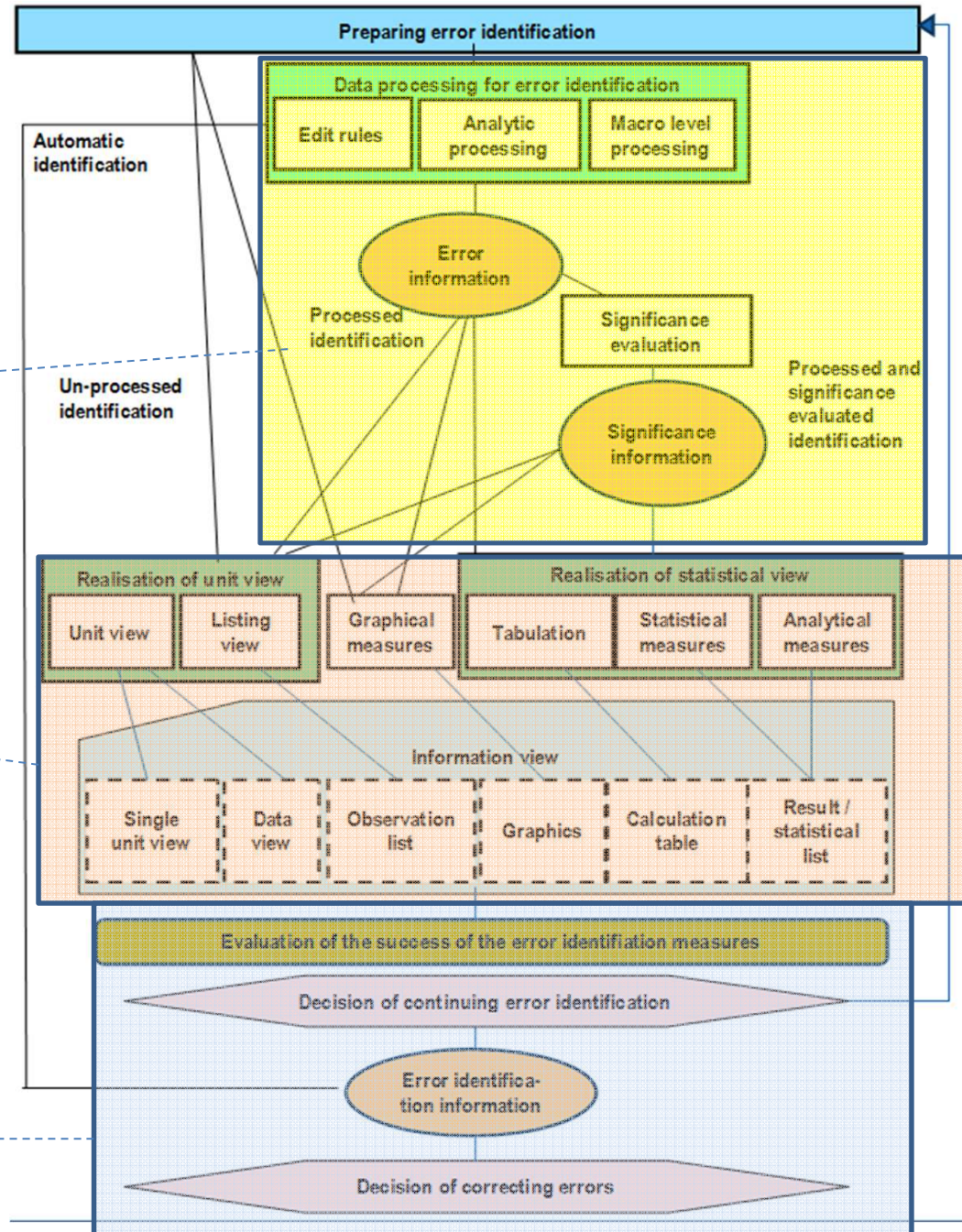
Editing process ERROR IDENTIFICATION

(probably subject to further specification)

Data processing

Realisation of information view for error identification

Evaluation of the error identification and decisions of further measures

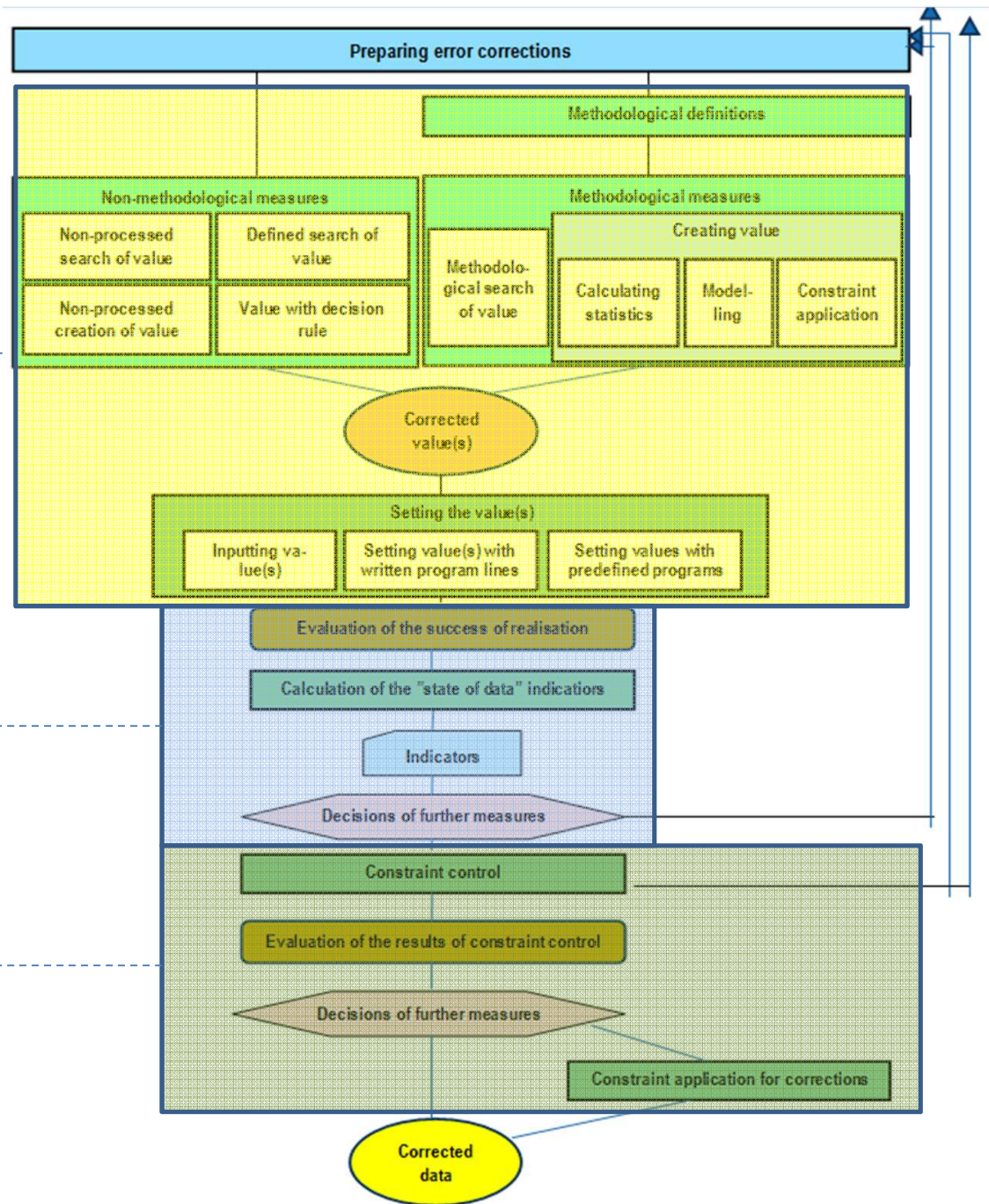


Editing process ERROR CORRECTION

Correction of identified errors

Evaluation of realisation of correction ;
calculation of "state of data" indicators

Controlling constraints of the edited data and possible corrections



Process and quality evaluation

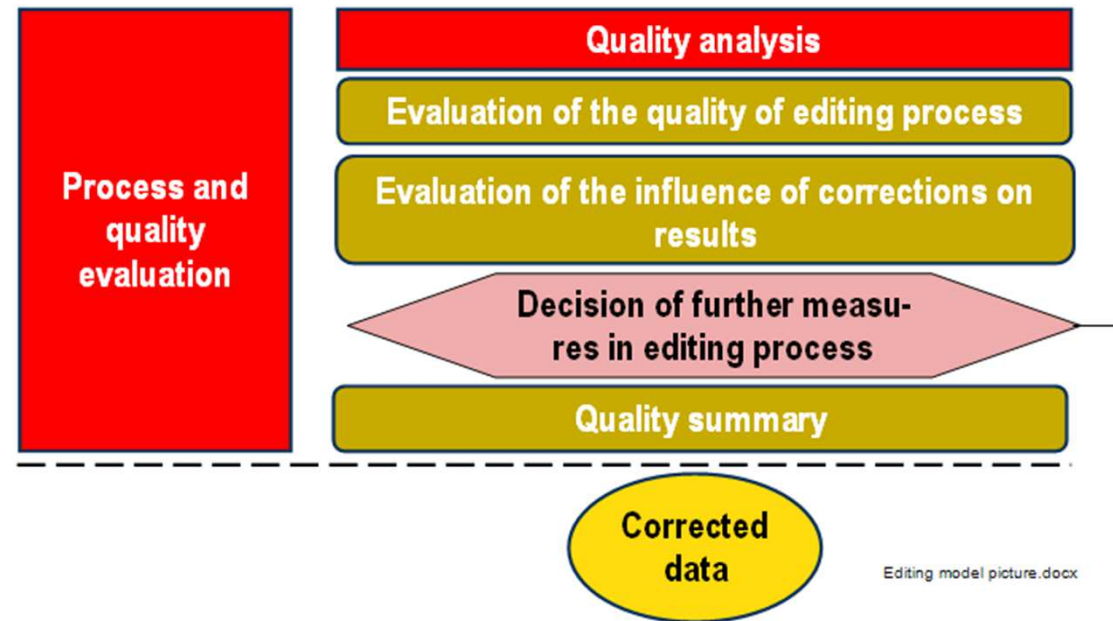
- Process and quality can be evaluated with indicators, which should be calculated automatically. The process of calculation is in a constant form.

- **Indicators describing the editing process**

- **"State of data" indicators** (essential estimates at the population level and in relevant subgroups, as in preliminary analysis and during editing process)

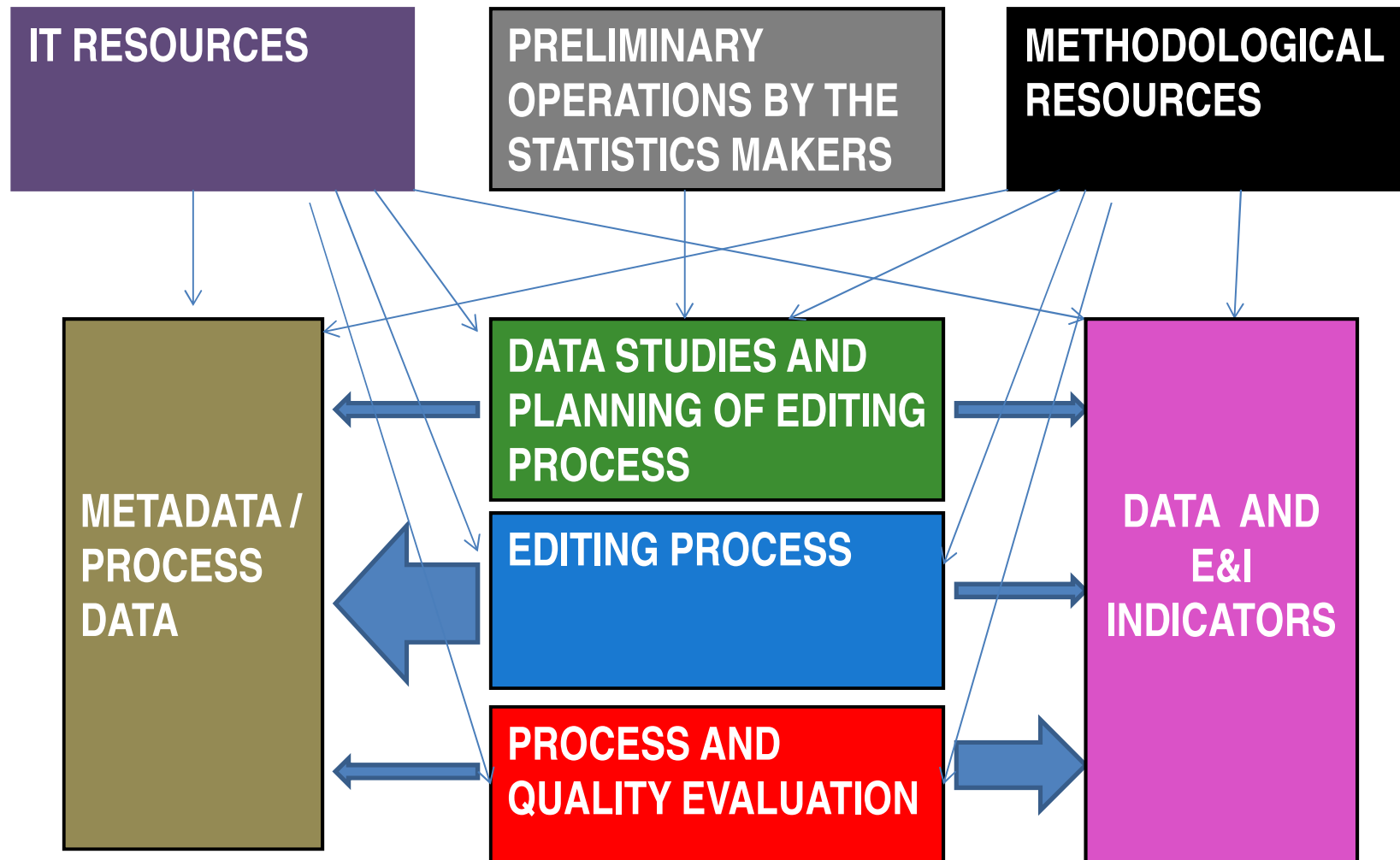
- **Indicators revealing the influence of editing on results**

- **Indicators in relation with previous results**



See Saara Oinonen's presentation and paper

3. Realisation of the process model: methodologies, practices, IT solutions



METHODOLOGICAL RESOURCES

Methodology bank

- The actions realised in the editing model are supported with the knowledge included in the methodology bank, which describes the methods included in the methodology groups in the different phases of the editing model.
- **Method** as a term can be considered here broadly: in addition to *statistical*, *mathematical* and *logical actions* it includes *consistent courses of actions*.
- The structure of the methodology bank follows strictly the **methodology groups** appearing in the editing model.

Measures describing the data	Refining the data	Search of value	Creating value
Realisation of unit view Realisation of listing view Calculation of statistical measures Realisation of tabulation Realisation of analytical measures Realisation of graphics	Edit rules Analytic processing Macro level processing Significance evaluation	Non-processed search of value Defined search of value Methodological search of value	Non-processing creation of value Value with decision rule Value with calculating statistics Value with modelling Value with constraint application
		Setting the value Inputting value Setting values with written program lines Values with predefined programs	

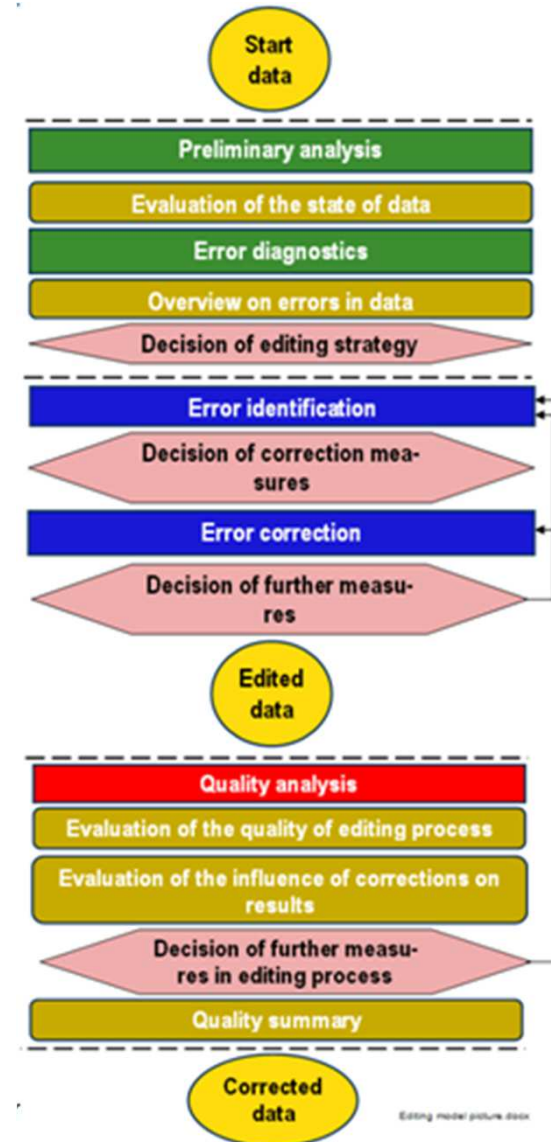
Concept library

- **Concept library** defines the concepts used in the model and the methodology bank
- Methodology bank and concept library should be easily available whenever needed (e.g. wiki-based). These could be utilised for the documentation of the quality of editing.

METHODOLOGICAL RESOURCES

Instructions for actions at different phases

- In the process model of editing statistics makers do decisions about actions to be done in different phases. The decisions are based on previous information and current data based information. The actions are carried out with chosen methods taking into account the characters of the data.
- For **decisions** (forthcoming actions and choices of methods), and for **interpretations** and **evaluations** of the results gained before and during processes and for **actions** required for realizing the methods there must be instruction collection, which helps during different phases.
- The instruction collection is based on research work and recommendations, international experiences and practices and Statistical office's experiences on data sets, error types and practices.



IT RESOURCES

IT solutions

- 1) IT environment should provide solutions to methods existing in the methodology bank (e.g. modules, procedures, macro packages) or at least a flexible platform to construct a program or other ways of action for the method. For larger entities of methods and practices it may provide applications or systems. The environment can include existing software (e.g. Banff, Selekt, LogiPlus) for the realization of some parts.
- 2) IT environment should allow flexible processing and obtaining of metadata and process data in order to control the process (E&I indicators) and the state of data (data indicators) during process and evaluating the quality of the final data.

METADATA /
PROCESS
DATA

DATA STUDIES AND
PLANNING OF EDITING
PROCESS

EDITING PROCESS

PROCESS AND
QUALITY EVALUATION

DATA AND
E&I
INDICATORS

FROM

- User's programming
- Manual written updates of programs

TO

- Use of modules, procedures, macros, applications
- Case specific information, modification information, methodological choices with **parameterization**



BANFF (Statistics Canada)

PROJECT FORM in SAS EG

ADD-IN in SAS EG

Banff: Error Localisation for G:\kaltuo\kalaba2012.sas7bdat

Data

Options
Edits
Output

Data source: G:\kaltuo\kalaba2012.sas7bdat
Task filter: None

Sort Current Input Data

Columns to assign:

Name
KohdeNo
YritNimi
Yritsoite
YritNo
YritPt
YritKunta
YritNimi
TOL2002
YritAloitv
YritKieli
Vuosi
Vuosi

Column roles:

ID (Limit: 1)
<column required>
BY

Enables you to assign variables (fields in input data sets) to various roles.

Preview Code Run Save Cancel Help

Must assign one variable to the role 'ID'.

SAS Enterprise Guide - Banff Tutorial (Completed).egp

File Edit View Tasks Program Tools Help

Project Tree

- Project Process Flow
- Edits Verification
- historical
- current
- Edits Summary Statistics
- Query1 for current
- DATA1
- DATA2
- DATA3
- DATA4
- DATA5
- DATA6
- SASUSER.OUTLIER_INDATA
- Programs
- SASUSER.ERRORLOC_INDATA
- SASUSER.DETERMINISTIC_INDATA
- SASUSER.DONDR_INDATA
- Donor Imputation
- SASUSER.DONDR_INDATA2
- SASUSER.ESTIMATOR_INDATA
- Imputation by Estimation
- SASUSER.ESTIMATOR_INDATA2
- SASUSER.PROPRATE_INDATA
- SASUSER.CURRENT_DATA
- SASUSER.MASSIMPUTATION_INDATA
- Mass Imputation

Project Process Flow

Run Stop Export Schedule Zoom Project Log

Prepare data for ... SASUSER.D... Determini... Imputation output data reco... output status fo...

Server List

Refresh Disconnect Stop

Servers

Private OLAP Servers

Ready

Käynnistä Olilla Pauli... Posti - [K... Microsoft ... 2 SAS E... T

BANFF (Statistics Canada)

PROCEDURE FORM in SAS BASE / EG

BANFF PROCESSOR

jobid	seqno	process	specid	editgroupid	byid	acceptnegative
j1	1	verifiedits	spec01	eg1		y
j1	2	editstats		eg2	v1	y
j1	3	beforeoutlier				
j1	4	outlier	spec01		v1	n
j1	6	outlier	spec02		v1	y
j1	8	afteroutlier				
j1	9	errorloc	spec01	eg2	v1	y
j1	11	deterministic		eg2	v1	y
j1	14	donorimputation	spec01	eg2	v1	y
j1	17	donorimputation	spec02	eg2		y
j1	20	estimator	spec01		v1	n
j1	23	estimator	spec02		v1	y
j1	26	beforeprorate				
j1	27	prorate	spec01	eg6	v1	n
j1	30	prorate	spec02	eg7	v1	y
j1	31	afterProrate	macvar1			
j1	34	errorloc	spec02	eg2	v1	y
j1	36	beforeMassImputation				
j1	37	massimputation	spec01		v1	y

estimatorid	seqno	fieldid	auxvariables	algorithmname
est1	1	egg_laied	HEN_GE20	HISTRATIO
est1	2	hen_tot	EGG_LAID,HEN_TOT	HISTREG2
est1	3	hen_ge20	EGG_LAID,HEN_TOT	CURREG2
est1	4	egg_sold	EGG_LAID	EGGREG
est1	5	egg_valu	EGG_SOLD	CURREG
est1	6	hen_lt20		PREVALUE
est1	7	hen_oth	HEN_GE20	HISTRATIO
est2	1	qr_rev	HEN_GE20,EGG_SOLD,EGG_VALU	CURREG3
est2	2	qr_exp	QR_REV,HEN_TOT	CURREG2

```
%banffProcessor(
  keyVar=IDENT,
  dataLib=mylib,
  metaPath=C:\Metadata,
  jobId=j1,
  curFile=currdata,
  histFile=histdata,
  custProgFref=myprogs,
  logType=3,
  estimatorOutputType=3,
  seed=1
);
```

```
Data: G:\Editointi\BANFF\Tutorial\current.sas7bdat
Server: Local
```

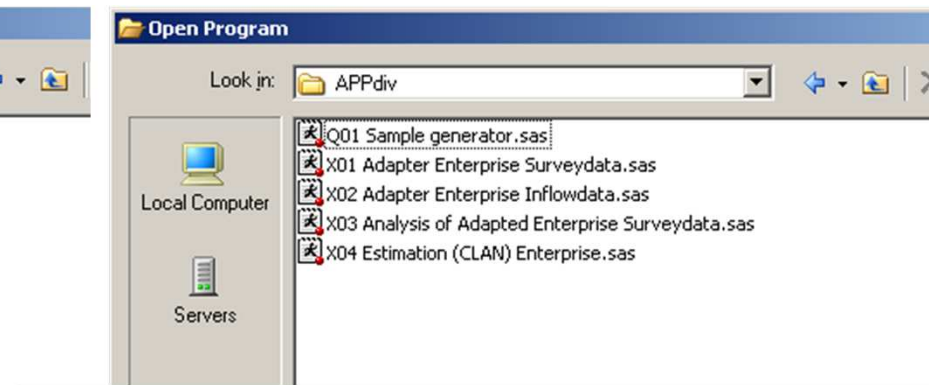
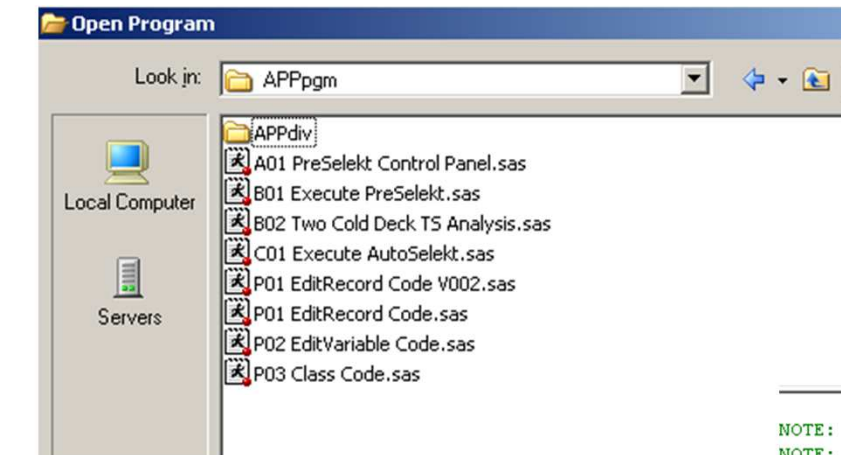
```
----- */
%_eg_conditional_dropds(sasuser.sta_outreducededits);
%_eg_conditional_dropds(sasuser.sta_outeditstatus);
%_eg_conditional_dropds(sasuser.sta_outkeditstatus);
%_eg_conditional_dropds(sasuser.sta_outglobalstatus);
%_eg_conditional_dropds(sasuser.sta_outeditapplic);
%_eg_conditional_dropds(sasuser.sta_outstatus);
```

```
----- */
/* Call PROC EDITSTATS to find Edits Summary Statistics.
```

```
PROC EDITSTATS
  DATA = ECLIB000.CURRENT
  OUTREDUCEDEDITS = sasuser.sta_outreducededits
  OUTEDITSTATUS = sasuser.sta_outeditstatus
  OUTKEDITSTATUS = sasuser.sta_outkeditstatus
  OUTGLOBALSTATUS = sasuser.sta_outglobalstatus
  OUTEDITAPPLIC = sasuser.sta_outeditapplic
  OUTVARSROLE = sasuser.sta_outstatus
  EDITS = "HEN_LT20 + HEN_GE20 + HEN_OTH = HEN_TOT;
2*EGG_LAID <= HEN_GE20;
HEN_GE20 <= 4*EGG_LAID;
EGG_SOLD <= EGG_LAID;
EGG_VALU <= 2.75*EGG_SOLD;
0.9*EGG_SOLD <= EGG_VALU;
HEN_GE20 <= 300000;
4 * EGG_VALU <= 1700000;
HEN_TOT >= EGG_LAID;
QR_REV = QR_EXP + QR_PROF;
0.5*QR_REV <= 23*EGG_VALU;
23*EGG_VALU <= 1.5*QR_REV;
HEN_LT20 >= 0;
HEN_GE20 >= 0;
HEN_OTH >= 0;
HEN_TOT >= 0;
EGG_LAID >= 0;
EGG_SOLD >= 0;
EGG_VALU >= 0;
QR_REV >= 0;
QR_EXP >= 0;"
  ;
  BY AREA;
RUN;
```

SELEKT *(Statistics Sweden)*

Based on very sophisticated and advanced macro system



Revised code: Anders Kraftling, Anders Norberg
Date: 2010-02-02

Changes:

```

NOTE: UNE_Var      [UNE_Y1]          First_Hours
NOTE: EDI_Var      [EDI_Y1]          Last_Hours
NOTE: EDI_Est      [ESTIM_Y1]        T_Hour
NOTE: UNE_Aux      [UNE_X1]          First_Time
NOTE: EDI_Aux      [EDI_X1]          Last_ATime
NOTE:               [_Y_max]         1 (Number of Y-vars)
=====
/*----- %Define_Survey UNEDITED survey variable name
[EDITED survey variable name] [optional]
[ESTIM variable for corresponding estimate of sum (with an initial p)] [optional]
[AUXILIARY variable to be used for UNEDITED data] [optional]
[AUXILIARY variable to be used for EDITED data] [optional]
The default value for Auxiliary variables=1 if omitted.
-----*/
Flag Consequence
-----
2 Flag single variables by a traditional edit check.
The suspicion is generalised to allow for any positive suspicion,
used in the computation of local scores for the corresponding survey
variables in selective editing.
----- %Define_Test UNEDITED test variable or expression
[EDITED test variable or expression] [optional]
-----*/

if not (0.8 < First_PopEmp/Frame_PopEmp < 1.25) and Susp_Y1<0.85 then do; /* Clean all definitions/mapping for Survey variables and Test expressions */
  ErrCode_106=2; Susp_Y1=0.85;
end;

if not (0.67 < First_PopEmp/Frame_PopEmp < 1.5) then do;
  ErrCode_107=2; Susp_Y1=1;
end;

if not (4000 < First_Turnover/First_PopEmp < 80000) then do;
  ErrCode_108a=2; if Susp_Y1<0.4 then Susp_Y1=0.4; if Susp_Y2<0.8 then ;
end;

if not (400 < First_Turnover/First_PopEmp < 800000) then do;
  ErrCode_108b=2; Susp_Y2=0.6; Susp_Y2=0.95;
end;

/* Survey variable #1 */
%Define_Survey(First_PopEmp,Last_PopEmp,Tot_PopEmp,1,1);
/* Test variables for survey variable #1 */
%Define_Test(First_PopEmp/Frame_PopEmp,Last_PopEmp/Frame_PopEmp);
%Define_Test(First_PopEmp/First_Turnover,Last_PopEmp/First_Turnover);

/* Survey variable #2 */
%Define_Survey(First_Turnover,Last_Turnover,Tot_Turnover);
/* Test variables for survey variable #2 */
%Define_Test(First_Turnover/Pre_Turnover,Last_Turnover/Pre_Turnover);

```

Preparations for Selekt (Statistics Finland)

```
%let Ad_year = 2009;
%let Ad_quarter = 0; * If the data is at the year level, put 0, otherwise quarter / month;
%let Unedited = Asunto09 ; * Name of unedited data;
%let Edited = Aineisto_valmis_09 ; * Name of edited data;
%let Sample = Otos2009 ; * Name of sample data;
%let Frame = ; * Name of frame data;

options ls=120 notes errors=3;

*****;
* PO 7.8.2012;
* Possibilities for modifying unedited data, edited data, sample data, frame data;
* Sample data or frame data are used only for design calculations;
* Modification macros are lines for subsequent data phases;
*****;

%macro Unedited_Modification ;
    _year_ = %ad_year; _quarter_=%ad_quarter ; _dummy_=1;
    _id_=ltun; * eri id-nimi!;
    * Pieni nimieroavaisuus, korjataan sijoituksella;
    asunnotpa_oh=asuntopa_oh;
    asunnotpa_to=asuntopa_to;
%mend Unedited_Modification ;

%macro Edited_Modification ;
    _year_ = %ad_year; _quarter_=%ad_quarter ; _dummy_=1;
    _id_=ltunn;
    Resp_%ad_year._%ad_quarter = 1;
    * Pieni nimieroavaisuus, korjataan sijoituksella;
    asunnotpa_oh=asuntopa_oh;
    asunnotpa_to=asuntopa_to;
%mend Edited_Modification ;

%macro Sample_Modification ;
    _year_ = %ad_year; _quarter_=%ad_quarter ; _dummy_=1;
    _id_=ltunn;
    Sample_%ad_year._%ad_quarter = 1;
    rename alue=alue_osite; * Varsinaisessa tiedostossa alue-mu
%mend Sample_Modification ;

*****;
* PO 7.8.2012;
* Realisation of modifications. The target is to harmonise
  the variables to be used in the SELEKT phases;
*****;
* Unedited data processing;
%Variable_Harmonisation(%unedited ,unedited,%ad_year ,%ad_quarter ,Unedited_Modification);
* Edited data processing;
%Variable_Harmonisation(%edited ,edited,%ad_year ,%ad_quarter ,Edited_Modification);
* Sample (including nonresponse and overcoverage) processing;
%Variable_Harmonisation(%sample ,sample,%ad_year ,%ad_quarter ,Sample_Modification);
*****;
* PO 7.8.2012;
* Size of stratum, number of respondents and final stratum;
* Alternatives: Design_Calculation_Via_Weights and Design_Calculation_Via_Frame ;
*****;
* TÄLLE TÄYTYY TEHDÄ VIELÄ ERI TAPAUKSIEN TOIMIVUUDEN TESTAUS;
%Design_Calculation_Via_Weights(sample,edited,%ad_year ,%ad_quarter ,SamplingWeight,alue_osite ala_osite tikaryh);
*****;
* PO 7.8.2012;
* Previous values for next year calculated here (overall survey 0 or quarter/month 1,2, ...) ;
* To be transferred to next year via a permanent data set;
*****;
%Previous_Values(unedited,%ad_year ,%ad_quarter );
%Previous_Values(edited,%ad_year ,%ad_quarter );
```

```
*****;  
* A macro finds all existing Edited_, Unedited_, Sample_ and/or Frame_ data sets in the  
  defined library;  
* These data sets are used in constructing Survey_data_ and Inflow_data_ sets;  
* These are used in SELEKT but also for indicator calculations;  
* If the latest edited data exists as well, then it is included in the Current_survey_data_  
  set;  
*****;  
  
%Survey_Inflow_Creation;  
  
* Classes and other CLAN definitions are given in program 201 Variables _____.sas;  
  
%CLAN_Class_Creation;  
  
%CLAN_Calculation;  
  
%Estimate_Transfer;
```