PROCEDURES TO IMPROVE THE DATA CLEANING PROCESS BASED ON QUALITY INFORMATION

October 2003

UN/ECE WORK SESSION ON DATA EDITING

Overview

- Introduction
  - Quality framework
- Collection about information
- Example of Austrian Labour Force Survey
  - Improvement project
- Management Aspects
  - Conclusions from example
- Possible methods for evaluation

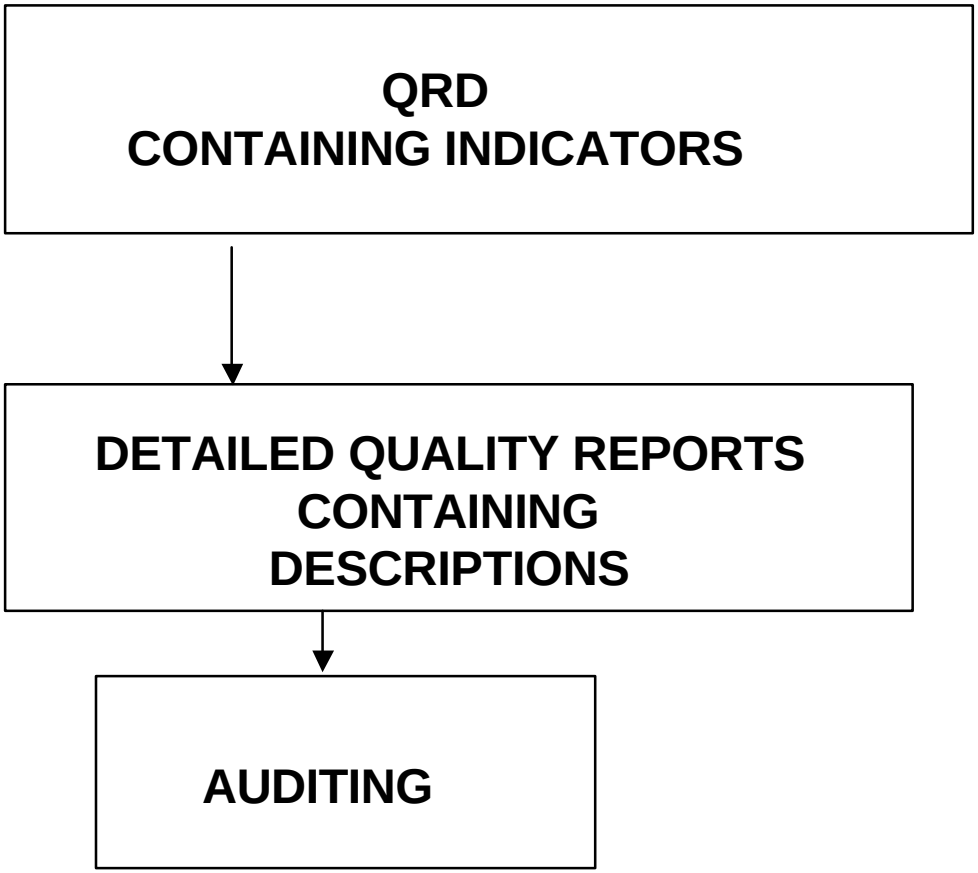
Metadata

- Demand for metadata is increasing
  - Not only for producer but also for customers
- Statistical Council is key observer of statistical products in Austria
- Data cleaning as a core process must be understood
  - More information by users required

Quality Framework(I)

- Product Quality is one of the piles of TQM
- Necessity to build up a quality reporting system
- Implementation during 2001 and 2002
  - QRD
  - Detailed Quality Reports
- First results and conclusions now available

Quality Framework(II)
Quality Report



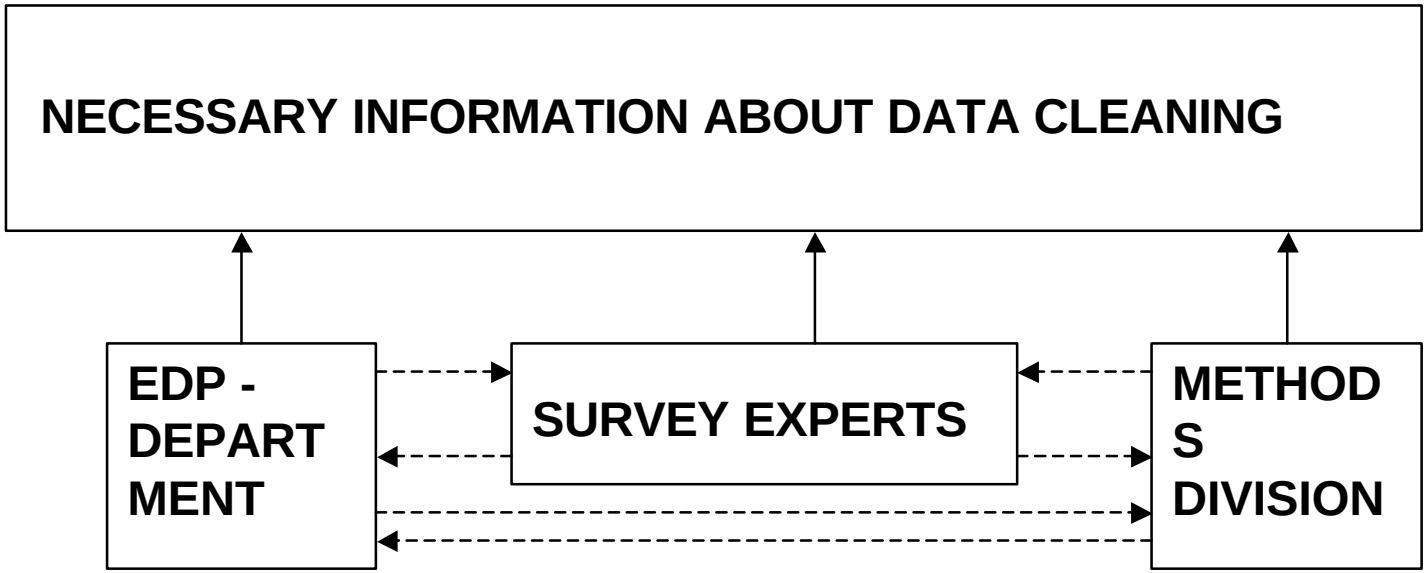
Indicators concerning Data Cleaning

- Indicators related to the data
  - Number of erroneous records
- Indicators about the process
  - Difficult to evaluate (analysis required)
  - Related to the management
  - Related to organization

Collecting information about data cleaning

- Information not always clear
  - Survey Manager not the one who implemented the procedures
- Not standardized information
- Information must be transferred in a usable form

Information flow



Problems when collecting information

- Not only one person has the whole information
- Often hidden sometimes even vanishing knowledge

First consequences

- Big improvement potential
- Deeper analysis of the data cleaning process
  - Increasing of academic staff
  - Demand on documentation
- Launch of improvement projects

Austrian Labour Force Survey

- Performed since 1995 in its current form
- Embedded in the Austrian Microcensus (quarterly sample survey 1% of the population)
- Microcensus has two parts
  - Basic program, mandatory
  - Special program, voluntary (in January of each year: LFS)

Non-Response in LFS

- Unit Non-Response
  - amounts 9-11%

- Item Non-Response
  - Complex questionnaire
  - Time consuming face to face Interview
  - Amounts up to 20%

Imputation (1995-2002) (I)

- EUROSTAT demanded complete data records
  - Imputation was necessary
- Based on information form the basic program, a distance based donor method was selected

Imputation (1995-2002) (II)

- Methods division received an order to develop a procedure for imputation
- Method was used as a black box by the survey experts
- Only one-dimensional checks of results were performed

Imputation New (I)

- In 2002 a detailed analysis of imputation process took place
  - Different parts of the LFS were investigated
  - Multidimensional tables
- Necessity of changing the imputation procedure
- Desire at survey staff to learn more about imputation methodology

Imputation new (II)

- Different process
  - Analysis
  - Consultation of methods
  - Selection of method (hot-deck)

- Stepwise procedure
  - Imputation was performed separately for different groups of variables

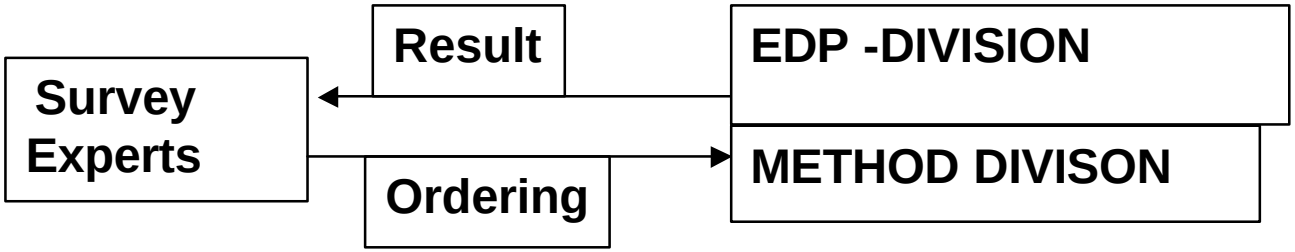
Quality Review

POSITIVE EFFECTS	NEGATIVE EFFECTS
ACCURACY strong COHERENCE partial CLARITY partial ACCESSIBILIT: A little	COMPARABILLITY Sometimes strong TIMELINESS once

Conclusions from LFS Example

- Organisatorical aspects are important
- Useful to have structure for an improvement project
- Transfer of knowledge to survey experts is necessary
- Project plan would have been helpful

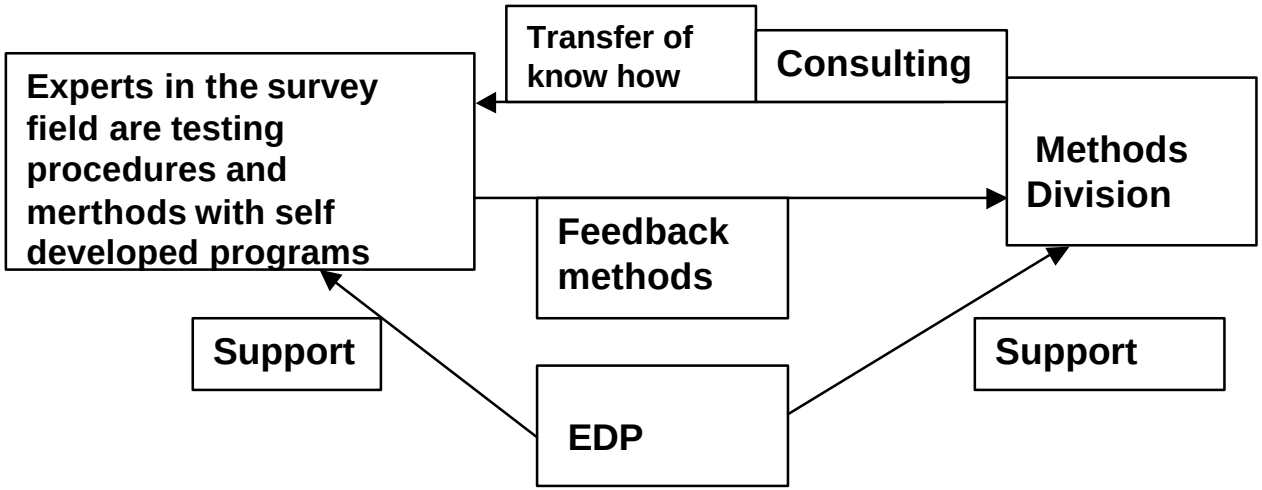
The old model(I)



The old model (II)

- Arrows are only unidirectional
- Knowledge concerning data cleaning is too centralized
- Methodologist lacks also on special knowledge

The new model (I)



The New Model(II)

- Methods and EDP consulting but not developing
- Knowledge transfer to survey experts
- All relevant knowledge is united so that questions from users can be answered more efficiently

Prerequisites

- Qualification of staff
  - Not only academic but trained in house
- Motivation from staff
  - Desire must come from survey experts
  - Job enrichment
- Support by high level management
  - user demands

Project plan for improvement of data cleaning

- Milestones are very important
  - Time consuming
- Written project plan
  - Why are you doing it
  - What are the goals

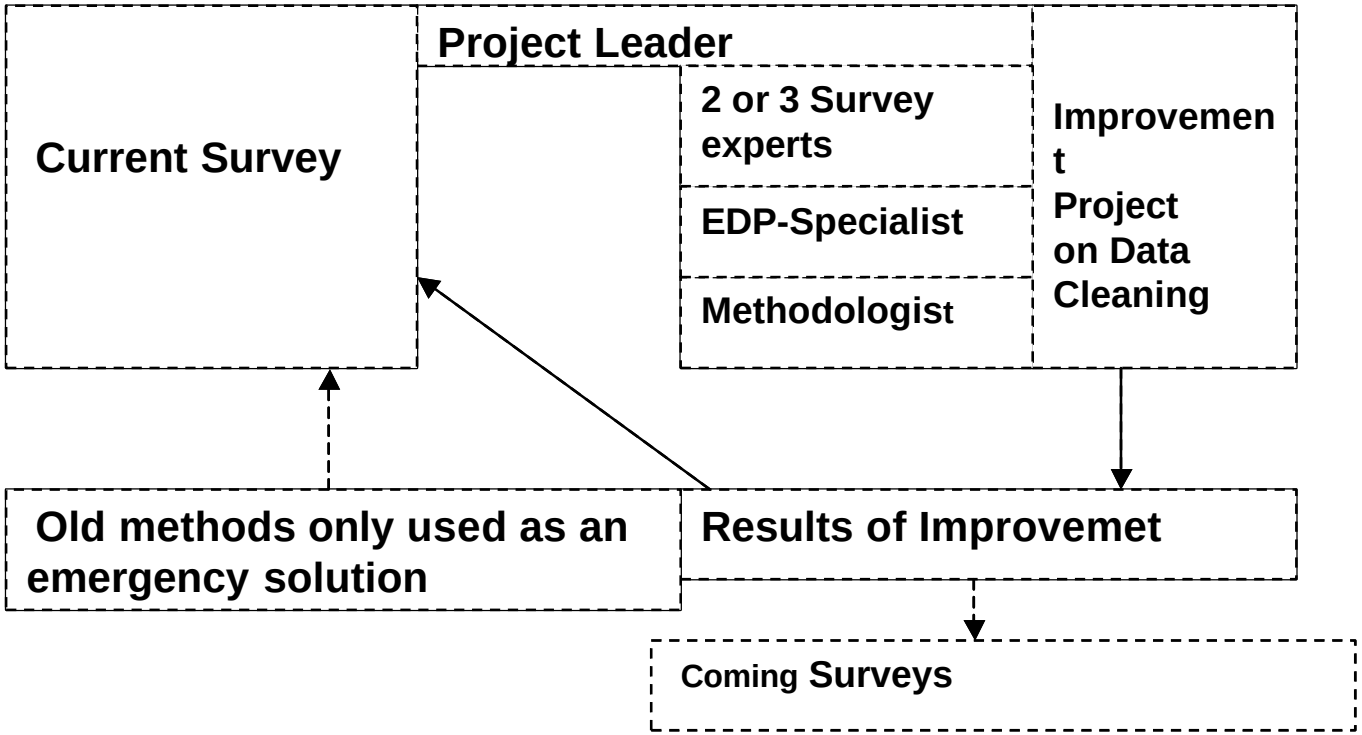
Project plan

1 .Nomination of Project Team
    -Distribution of tasks
2. Analysing of the actual situation in the data cleaning process
3. Discussion of new methods
    - What is state of the art, Study of methods used elsewhere
    - Consulting by methods division
    - Selection of suitable methods
    - Education of staff
4. Implementation of new method
    - Decision about software
    - Tests of results
5. Documentation of new methodology
    -Decision of publication strategy

Project Team

- Should not be that large
- Project Leader should be high in hierarchy
- Methodologist
- EDP-Specialist
- 2 or 3 experts from the subject matter department

Structure of the improvement project



Evaluation of Data Cleaning

- Decomposing the Quality of data cleaning
  - Organisational Aspects
  - Technical aspects
  - Quality of Data

Checklist for Evaluation

- MANAGEMENT AND ORGANISATION
  - Are the Methods of Data Cleaning well known in your division?
  - How many people have sound knowledge about the Data Cleaning in your division?
  - Are your methods approved by the methods division?
  - Do you have contact with other offices/organisations and compare your methods with theirs?
  - When did you perform your last improvement project?
  - Is your Data Cleaning Process fully documented?

- TECHNICAL ASPECTS
  - Is your Data Cleaning process fully automated?
  - Who developed the programs which run the data cleaning process?
  - How much support did you need from the EDP?

- DATA AND RESULTS
  - When did you perform your last ex-post study to evaluate the accuracy of the cleaned values?
  - Do you know on the effect your data cleaning has on the variance of your estimators?
  - Did you test your methods with a simulation study?

Plans

- Find potential for further improvement projects during feedback discussions
- Introduce new management model
- Develop detailed checklist for Data Cleaning
  - DESAP