I. INTRODUCTION

1. Statisticians face today different developments: increasing user demands especially regarding the timeliness of statistical results and the dissemination of anonymous microdata raise the demands on national statistical offices to overcome the trade-off between timeliness and accuracy with decreasing resources available. Another releasing and ongoing development is that more and more information is already available in (partly) pre-checked data. As there is greater worldwide competition, respondents push statistical offices to reduce the burden induced by surveys. Thus pre-checked data (from accounting systems) are more and more used and adapted to specific statistical needs. The closure of "gaps" in data or the deduction of reliable statistics from external data replaces more and more classical tasks of error detection and data correction.

2. Against this background, a joint working group of Destatis, the Federal Statistical Office (FSO) and the statistical offices of the Länder developed a new data editing concept. The FSO implemented the new concept in July 2004 through the introduction of materials, software, and IT modules. They support different activities related to (the planning of) data editing. The new software is embedded in a concept that is designed to strengthen the transmission of user demands on statistical results into an adequate and efficient data editing strategy. Efficiency will also be promoted by the standardization of data editing sub-processes which is an important precondition for the reuse of available methods and respective IT tools. A last important aspect is that the new IT tools support the manipulation of metadata as well as the editing of data.

3. Error prevention is another very important aspect of the new concept but not discussed in this paper. German official statistics boost their efforts to standardize and improve (electronic) questionnaires. In addition to these developments it haunts different approaches: enterprises may use Internet questionnaires or transmit data in predefined formats. The latest initiative of this approach is CORE, which provides the transmission tool "CORE.Reporter" free-of-charge to enterprises. It supports under Windows, Linux and MacOS data deliveries from enterprise specific accounting systems to one statistical portal. A second approach is the co-operation with external software vendors who implement

---

1 Prepared by Elmar Wein (elmar.wein@destatis.de).
2 CORE: Common Online Rawdata Entry
the contents of questionnaires in their business software such that business data are transferred to satisfy statistical needs.

4. The aim of this contribution is to give an overview of the new IT tools. First, new software for the planning of data editing will be presented in section II. It completes earlier SDE contributions on the planning of data editing. Section III contains information on new IT modules used for data editing. Special emphasis will be placed on the concept that serves as a frame for them.

II. IT SUPPORT FOR THE PLANNING OF DATA EDITING

The Concept of the Planning of Data Editing

5. The process for the planning of data editing was described in several contributions at the last data editing work sessions. The adapted project management methods subdivide the planning of data editing in the phases shown in the following figure. Grey marked phases indicate an IT support.

Figure 1: Phases of the planning of data editing

6. The planning of data editing consists of four main steps:
   - The collection and judgement of relevant information should ensure that all relevant information – especially on user demands – are collected and judged. At the end statisticians should possess considerations as regards a data editing strategy.
   - The purpose of the overall planning is the decision on a data editing strategy.
   - The aim of the detailed planning is to adapt and complete the data editing strategy.
   - The description of risks and the final check should verify and improve the preceding planning steps.

7. Figure 1 shows that most of the planning steps are supported by new IT tools. There is no need for an IT support of the phases "Description of Risks" and "Final Check". They are supported by methodological considerations, e.g. a checklist that contains relevant aspects of an optimal data editing strategy and possible risks. An IT support of the cost planning is only partly realised because of the different preconditions and accounting systems of the FSO and the statistical offices of the Länder.

IT Tools for the Planning of Data Editing

8. The IT tools mentioned in the following sections refer to a Data Editing Intranet site which will not described in this contribution. It serves like a manual for data editing and is disseminated among all German statistical offices.
Collection and Judgement of Relevant Information

9. An appropriate collection and judgement of relevant information lays the fundament of an output-oriented data editing strategy. First subject matter statisticians answer questions concerning the demand for statistical results, the possible beginning of data editing, and needed data from other statistics or statistical offices. Questions on available pre-information from pre-tests and pilot studies or other/similar surveys follow. They are supplemented by judgements of the survey contents as regards their possibilities to be checked. Specific questions of secondary statistics cover aspects like possibilities for pre-checking the data by data delivers. Attention will be also given to the aspects of the data collection process like the use of electronic questionnaires and the integration of checks. Questions on the available experience of the personnel and possible IT restrictions complete the catalogue of questions. Subject matter units are requested to describe possible risks and advantages of a data editing strategy to be planned. Finally they are requested to make suggestions for a confinement of the risks with the aim that a balance between the aspects mentioned in this section will be achieved. It is expected that IT production managers develop first considerations on a data editing strategy when they read the answers of subject matter statisticians.

10. All these steps are supported by a database that has been developed on the basis of ACCESS.\(^3\) Reports form the output of the database. They promote discussions and sharing of knowledge and experience. The database will be provided as a standalone application but its underlying relational data model is so powerful that it may be also implemented as a central database in a decentralised statistical system. The main reason for the choice of a database is the planned reuse of information for subsequent planning steps. It is planned to enhance the database by the integration of the data emerging from the overall and detailed planning. Another important aspect is that the planning information is linked to the classification of all German federal statistics (input-output-relation).

11. We expect the following benefits from the electronic guideline:
- A better support of the decision on a data editing strategy that bears in mind all relevant aspects – especially users' demands on statistical results.
- An easier documentation of available information, judgements, and considerations so that other statisticians can easily verify them and reuse them for subsequent similar planning steps for other surveys (preservation of knowledge and experience).
- A better support of less experienced subject matter statisticians to perform a systematic planning of data editing.

Overall Planning on the Basis of Predefined Process Chains

12. The aim of the overall planning is the development of a data editing strategy that means a sequence of data editing sub processes. The relevant information for the choice of a given process chain is provided by the electronic guideline mentioned above. There are three process sequences which consist of optional and mandatory data editing sub processes such that they can be adapted to survey specific needs. Two modular process chains have been developed for data editing as part of the data gathering process and one sequence for data editing by subject matter statisticians.

13. The process sequences for data editing as part of the data gathering process start with the dissemination of electronic / paper questionnaires or occurrence of data in enterprises / external offices and terminate with the provision of (partly) plausible data for data editing as part of data processing. The sequences cover the employment of paper and Internet questionnaires as well as the accruement of data in enterprises and external offices.

14. Data editing as part of the data processing is determined by the provision of demanded statistical results. It is assumed that a statistical office may disseminate up to four different types of statistical results: tendencies, preliminary results, standard results and micro data. They differ as regards timeliness and accuracy, their used data editing methods and survey contents. A data editing strategy, that means

---

\(^3\) The author thanks Tatjana Theis and Volker Stutzer for the development of the electronic guideline.
one data editing process chain corresponds to a type of statistical result. Subsequent data editing processes (=data editing strategies) complement each another. The entire process chain is rather complex because it documents all possible data editing strategies. It begins with preparing activities (optional sub process) and terminates with the mandatory provision of (partly) plausible (meta) data for subsequent analysis / tabulation. It is expected that it is so flexible that it will support the great majority of needed data editing strategies, and it represents best practices. Data editing methods may contain subject matter oriented checks and combinations of selective editing methods and automatic edit and imputation methods.

15. It is assumed that the use of more powerful data editing methods and IT tools and the ongoing budget cuts will reinforce the specialisation in the statistical offices. Consequently, the IT production manager will plan the data editing strategies and manage the data processing. Given the information from the electronic guideline the IT production manager chooses the data editing strategy (pre-defined process chain) that fulfils subject matter demands best. Adaptations of a given data editing strategy may take place where needed so that the final process sequence serve as a frame for the following detailed planning of data editing.

16. The process chains were presented for the first time at a training in September 2004. The subject matter statisticians mentioned that the chains clearly show the necessary steps of a data processing and seem to be a good base for survey specific data editing strategies. The IT production managers also appraise them positive but argued that they should be completed by the respective plans of data flows. This initiative will strengthen the approach. We expect the following benefits from the use of the process sequences:

- They serve as a frame and promote the development of best practices for data editing with positive effects on the efficiency and comparability of data editing processes.
- They facilitate a standardisation of data editing processes, methods, and IT tools. We hope that we can relieve scarce IT capacities by the standardisation of data editing processes so that there are better preconditions for the development of selective data editing and automatic edit and imputation methods. Another benefit of standardised IT processes may be better preconditions for a flexible employment of personnel.
- The process chains facilitate the detailed planning of data editing.

Specifying Checks with the PL-Editor

17. Specifying checks represents the beginning of the detailed planning. It requires a data editing strategy planned during the precedent overall planning because the edit are linked to data editing sub processes. The specifying process was redesigned as follows:

- Subject matter statisticians specify the survey contents. An object-oriented approach enables a structure of the specifications close to reality, e.g. a specified characteristic represents an object that possesses a question and answer texts, an explanatory text, a field for one or more value(s), and a coding or range check. Interrelated characteristics can be combined in a content unit. It contains consistency checks, classifications, functions, properties and sequences of characteristics and checks, e.g. a sequence for the context "electronic questionnaire" and another one for the context "standard results". Content units may contain inferior content units. Characteristics and content units are stored in a central database so that they are accessible and may be reused by other subject matter units. Consequently, if a subject matter statistician refers to a specified characteristic from another subject matter unit all elements of the characteristic mentioned above are automatically used.
- The specification syntax has been completely renewed: special characters from the existing syntax were removed and it was expanded by common functions, e.g. "mean", "median". Self-explanatory names for characteristics and content units are supported.
- The IT tool is supplemented by methodological recommendations provided by the data editing intranet site. They propose a sequence of checks with completeness checks at the beginning and
checks between actual and previous surveys at the end. Special attention will be given to the integration of checks in electronic questionnaires.

18. Greater demands on the efficiency of the data editing processes and the modifications mentioned above led to the development of a new application, called the PL-Editor\textsuperscript{4} that is a client-server-application. All data editing specifications will be stored in an office wide database that will be mirrored among all statistical offices of Statistics Germany. The following figure shows the user interface of the new IT tool which was designed in the well known Windows style so that the users can easily find functions which they needed:

Figure 2: User interface of the PL-Editor

19. Besides the reuse of existing specifications mentioned above the new IT tool provides the following functionalities:
- It generates executable IT modules in JAVA on the basis of edit specifications. The modules perform the checks and can be used for Internet questionnaires as well as for IT applications to be used for data editing by subject matter statisticians. The JAVA-Generator is the first module other generators can be integrated because the specifications are stored in a XML-format. This flexible approach enables the support of different IT platforms that are used for data editing. Other parts of the metadata, e.g. question and answer texts are used for the development of Internet questionnaires.
- In addition to a syntax checker an analysis component checks inconstancies between several specifications and displays them, e.g. correctly specified checks which are not referenced by existing sequences. The users can easily reach the respective elements for corrections via hyperlinks.
- Subject matter statisticians can create documentations, e.g. coding lists and the documentation of checks and characteristics. As the metadata for the documentations will be provided in the CSV-

---

\textsuperscript{4} The author thanks Andree Hähnel, Corina Teichmann, Alf Steilen, and colleagues for the development of the PL-Editor.
format they will also be able to use them for their own documentations. One of the next releases will contain a document editor.

- An online help-facility (in HTML) offers methodological recommendations and examples to be taken over in own specifications.
- The PL-Editor generates a file structure on the basis of specified characteristics and content units and displays the respective graphical documentations. The way back that means the integration of existing file structures in the object oriented database is also possible because they can be manually connected to new content units.
- The existing metadata of the specifications can be reused for tabulations with “STATSPEZ” (a tailor made application for complex tabulations). A transfer to SAS and their provision for statistical online databases is planned. The transfer of the metadata to other statistical software packages like SPSS is feasible because of the independent meta language and flexible integration of generators.

20. An important precondition for the reception of the benefits mentioned above is that subject matter statisticians comprehend the object oriented approach so that they are able to structure their specifications in accordance with their questionnaires. The pilot training held in September 2004 showed that the subject matter units can handle the new IT tool which effected positive reactions among the participants.

21. The PL-Editor was launched in July 2004 and is now used for 17 statistics. There are around 60 persons of the FSO to be trained in 2005. We expect the following benefits from its use:

- The implementation of the IT tool promotes the establishment of a well-structured metadata inventory.
- It will improve the efficiency of the specifying process, enables a multiple use of metadata in following survey processes, disburden subject matter units from the provision of metadata for management purposes and reduce the effort for the IT development.

Detailed Planning of Data Editing Processes with Process Managers

22. The planning of data editing processes forms the end of the detailed planning. It requires the planning steps mentioned above and starts with the estimation of the time effort for the processes of one chosen process sequence. The time effort is estimated by the methods derived from the manpower requirements. On the basis of the estimated time effort subject matter statisticians can assign personnel to the chosen data editing sub processes and estimate the costs for the personnel. At last the time scheduling can be done on the basis of the subsequent planning steps and bears in mind working hours per day and working days per week.

23. The planning activities led to the development of the data editing process managers which were realised with EXCEL 97 and work also with EXCEL 2002. The new IT tools set low demands on the hard- and software that is available in all statistical offices. We expect that the provided functionalities satisfy the demands of existing data editing strategies as well as the ones in the future. So we see no need for the use of sophisticated project management software for the majority of our surveys. A process manager consists of seven work sheets: general information on a survey and the planning, estimation of the time effort for processes, employment of personnel, time scheduling, estimation of the time effort of basic data editing activities (optional), overview of data editing process sequences (as help), and a work sheet with explanations. Both data editing process managers are linked to the three pre-defined data process sequences which are described in section 0 "Overall Planning on the Basis of Predefined Process Chains", page 3. Hyperlinks to the data editing intranet site lead to additional methodological help. The implemented macros support comparisons between plans and practice such that the causes of differences become obvious, e.g. whether they are caused by deviations due to different numbers of records or additional time for checking and coding. The following figure shows the table for the planning of the personnel employment:

---
5 The author thanks Tatjana Theis for the development of the data editing process managers.
24. The new IT tools were presented in workshops and a pilot training. We expect the following benefits from their use:
   - The additional help for the estimation of the time effort for typical data editing processes combined with the use of accepted methods of personnel requirements will lead to more realistic estimations for data editing processes and thus improve the employment of scarce resources.
   - The comparisons between planned and realised data/costs will provide better information for the management of data editing processes and improve the time scheduling of Statistics Germany.

III. NEW IT TOOLS AND MATERIALS FOR DATA EDITING

Methodological Background

25. The IT tools for data editing are embedded in a concept which is based on the three standardised and modifiable sequences of data editing sub processes. It is assumed that they cover a considerable majority (> 80%) of all existing data editing strategies. The processes are subdivided in typical activities and adequate methods are assigned to them. They form a collection of data editing methods, (graphical) analysis methods, coding methods and general data manipulation methods. The collection of methods shall represent a tool box; the methods shall be reused so that the burdens of IT specialists and methodologist can be reduced. The collection is disseminated via the data editing site to all statistical offices. A description of a method contains an explanation, recommendations as regards a combination with other methods, advantages and limitations of a method and further literacy. Most of the descriptions are derived from literacy but they are more and more supplemented by practical experience. The first use of the collection by a subject matter unit shows that there is a need for more detailed explanations.

IT Support of a Selective and Macro Editing

26. The Federal Statistical Office tested successfully a combination of a simple selective and macro editing method in 2004. An important aspect of the situation was that preliminary, projected results on
costs of the producing industry have to be disseminated after eight weeks. As regards data editing the survey was partitioned by branches and distributed among subject matter statisticians. The situation requires a priority setting among erroneous records within a stratum. Thus it was decided to prioritise erroneous records of a stratum by a selective editing method and to prioritise strata (branches) by a macro editing method. Both methods were realised with SAS macros.

27. The macro which supports the selective editing can rank erroneous records by computing error scores on absolute differences between estimated plausible values and current (raw) values in relation to the sum of plausible values. The developed selective editing method can be parameterised in different ways:
- the estimated plausible value may be also a central metric (mean/median) of a current / previous sample,
- the Chebyshev, Minkowski or Mahalanobis distance may be used for error detection,
- the differences may be weighted by subject matter, methodological, and projection factors,
- error scores may be supplemented by survey specific indicators, e.g. accurateness indicators or subject matter oriented checks.

28. The macro editing method prioritises strata on the basis of the following factors:
- an indicator of possible plausibility improvements: as there are small strata we computed the range of the error scores per stratum divided by the median of the error scores,
- an "importance factor" of a stratum: we used the number of employees and the turnover of a stratum related to all employees and the total of turnover of the producing industry,
- a “comparability factor”: we compare the sums of "stable" characteristics of the current sample with the ones of the previous sample,
- a projection factor: all previous factors are based on the current and previous sample that means only ten of 16 000 records were used for the dissemination of preliminary projected results. Thus we decided to combine the previous mentioned factors with this one.

29. The test required the development of around 50 reusable and configurable SAS macros. SAS was used because it restricts the effort for the development, it enables the exchange of tools with other statistical offices and it is a standard-software of the Federal Statistical Office. The macros support the development, test, performance of the methods, and their ex post analysis. They relate to the collection of the methods used for data editing but most of them support basic activities e.g. checking of parameters. In general the macros can be categorised as follows:

Figure 4: Categorisation of SAS macros used for selective and macro editing

Around 25 macros support general tasks e.g. the computation of common descriptive statistics, the tabulation or the computation of weighting factors.

30. The macros used for the production of a statistic possess a unique structure and are documented in English. First they check the parameters, second they harmonise different input files (if necessary) then they perform the main manipulations and document key indicators of the operations in external files (HTML-, RTF-, PDF-Format). At the end a macro deletes all temporary files – if set by a parameter. They require Base SAS 8.2, SAS/Stat, and SAS/Graph. They are designed to be integrated via JAVA-classes in JAVA-applications for data editing and terminate with standardised code and messages. This approach will be tested in the first half of 2005 and an IT application will be realised for the production of statistics on costs of the building industry. The use of the same macros by data editing and methodologists is very efficient because it releases IT specialists.
31. The macros are used by seven standardised SAS projects and are supplemented by a guideline for the development, test, use, and post analysis of a selective editing method (in English). The SAS projects support the import of data in the SAS system, specific analysis (of errors), the development and computation of error scores, the simulation of a developed selective editing method, the computation of data for the use of a method for the production of statistics. Two SAS projects are realised for the provisional use of a selective and macro editing method besides an existing IT application for data editing. One project supports the editing methods and the other one supports comparisons between plausible data of a previous and current survey.

32. The great majority of the macros were developed for the support of enterprise surveys on costs. To increase their reusability they were parameterised to a great extent, possible demands were implemented, and the preconditions as regards the file structures and number of strata were lowered. This programming and testing was very complex and resource consuming. However when some macros were used for the survey on the European Intratrade it became obvious that the macro approach is still too restrictive because we had to develop another SAS macro for the survey specific selective editing method. Thus it was decided to enhance the flexibility of the SAS macros and the considerations led to the development of a macro template. A template is a macro that consists of the SAS macros mentioned above and may be supplemented by survey specific data steps. It supports a specified purpose e.g. the computation of error scores or aggregates of strata but it offers more alternatives of data manipulation. The first template was realised for the support of the macro editing method because it possesses a lot of options besides the factors mentioned in this paragraph.

IT Support of an Automatic Editing

33. The Federal Statistical Office developed a prototype of an error determination program on the basis of the methodology of the Dutch CBS. The approach was chosen because it can handle data structures which consist of categorical and metric characteristics – a constellation which seems to be typical for enterprise surveys. Another reason for the choice is the promising test results. The modules which determine possible erroneous characteristics and eliminate the influence of the respective fields on specified checks are developed in the programming language C. The prototype has been tested since December 2004.

34. The Federal Statistical Office will use IVEWare for automatic corrections. This software was chosen because of its numerous opportunities to build statistical models. Besides this software the SAS procedures MI, MI Analyse and a SAS implementation of the Fellegi-Holt-Methodology are available for automatic corrections. Due to their restricted facilities of model building (SAS procedures) and their longer run times (Fellegi-Holt-Methodology) they have not been used until today. Current research activities are also focussed on multiple imputations with artificial neural networks.

---


4 Further details on the data editing intranet site contains the contribution of Elmar Wein (2004): Improvement of Data Editing Processes, Q2004, Proceedings of the Conference

6 The author thanks Carsten Kuchler for the development of the prototype "missfix".
vi For further details of the method see Elmar Wein (2005): Introducing and Implementing a New Data Editing Strategy, SDE work session, Ottawa 2005