WP. 19

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (ii): Developments related to new methods and techniques

IT TOOLS FOR AN INTEGRATED DATA EDITING CONCEPT

Supporting Paper

Submitted by the Federal Statistical Office Germany[1]

I. INTRODUCTION

1. In 2002 the German Federal Statistical Office started to implement an integrated data editing concept, covering the planning and execution of the underlying work steps as well as the applied methods and algorithms. During the current phase of realisation, the emphasis lies on developing IT tools supporting the significant planning phases of an data editing process. This paper presents three prototypical IT tools that are already available or will be launched together with a preparatory training within the next months:

- DE reference: the intranet pages (section III),
- DE scheduling: collection and assessment of planning conditions (section IV),
- DE specification: the editor (section V).

Beside a brief summary of their functionality the presentation will primarily focus on the underlying models of the information flow between the sub processes and the context of application. In order to substantiate the implementation of the sub processes, the following section initially reveals the primarily application requirements for the IT tools.

II. APPLICATION REQUIREMENTS FOR THE IT TOOLS

2. A major objective during the modelling of the data editing processes was the idea of decentralisation. This means that the IT tools should rather enable the department concerned with a specific survey to plan and carry out data editing in their own responsibility than support a central methodological unit. This approach affects (a) the assumed skill level of the potential users and in this regard their guidance within the IT tools and (b) the question of how to document the work steps of a specific survey and make them available for subsequent surveys and other departments.

3. One serious problem arises from the linkage of the decentralisation approach and the period of the application flow of a specific survey. Since it may cover several months or even years, the staff of the responsible department is carrying out the data editing sub processes rather infrequent. To prevent permanent refreshment courses, the IT tools should therefore provide extended help and documentation

[1] Prepared by Carsten Kuchler, carsten.kuchler@destatis.de.

facilities comprising both a brief description of their functionality and a step-by-step workflow instruction. In addition there is a comprehensive reference needed that offers detailed descriptions of the sub processes and relates adjacent sub processes with regard to the superior data editing process.

4. Since every IT tool is dedicated to a specific sub process of data editing, the information flow between the sub processes and their associated IT tools demands flexible interfaces. Moreover one has to consider that the application flow of a data editing process may change due to added or suspended sub processes. Thus the declaration of the input/output streams between the IT tools needs to be independent from the potentially varying information flow between the sub processes. The requirement of flexible interfaces also affects the facility of supplying documentation and results to other parties as mentioned in the second paragraph.

III. DE REFERENCE: THE INTRANET PAGES

5. The DE intranet pages provide an online companion covering at least all topics concerned with the data editing process in the Federal Statistical Office. In particular they achieve two major aims: (a) supply a reference for the application flow and the available methods of the data editing sub processes and (b) offer a guidance for the planning and execution of the complete data editing process that relates single work steps to adjacent sub processes and thus provides a rather holistic view compared to the specialised help facilities of the IT tools. Since the implementation of the data editing concept was considered right from the start as a dynamic process, the decision for an online instead of a desktop companion was primarily motivated by the possibility to update the document at low cost and with a maximum of flexibility.

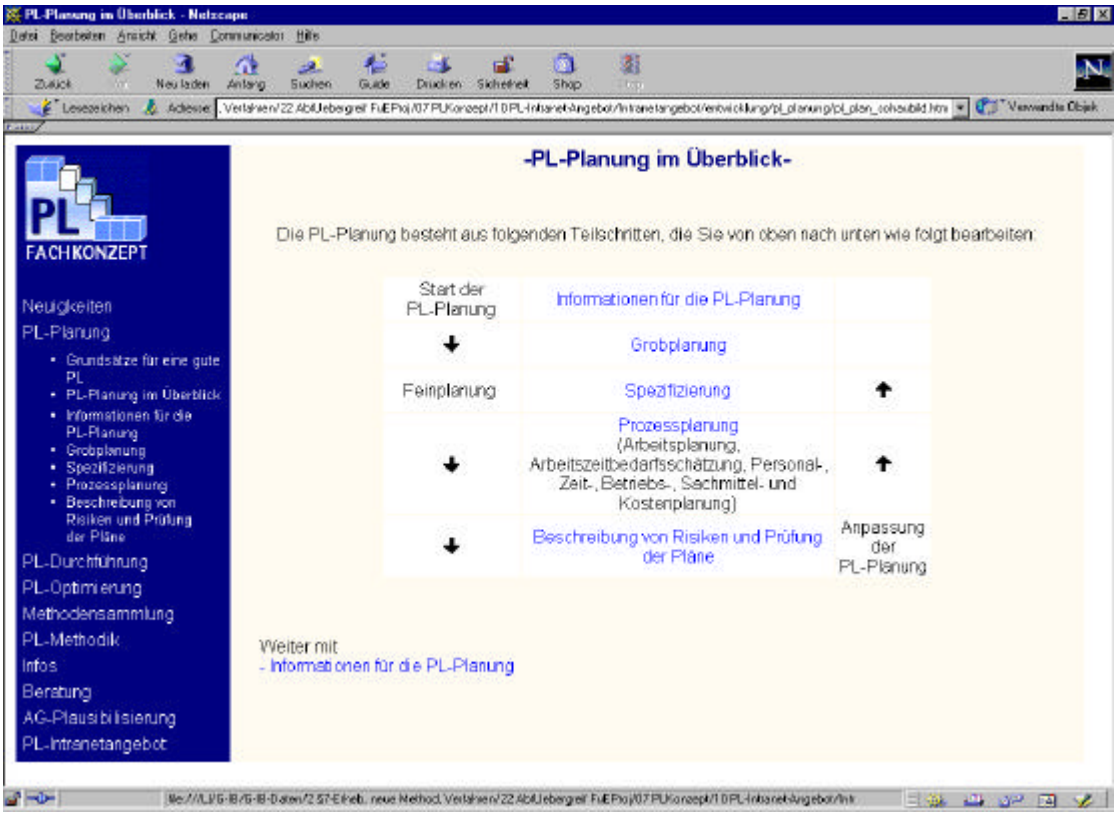

Figure 1: Overview of the sub processes concerning the planning of a data editing process. Blue highlighted words provide links to more detailed pages.

6. As an essential element the DE intranet pagers provide a complete collection of the data editing methods applicable during the execution of data editing processes with the associated IT tools. This collection serves both as a online reference and a definition of the mandatory status quo of data editing in the Statistical Office that is regularly updated in accordance with the users in the departments applying these methods.

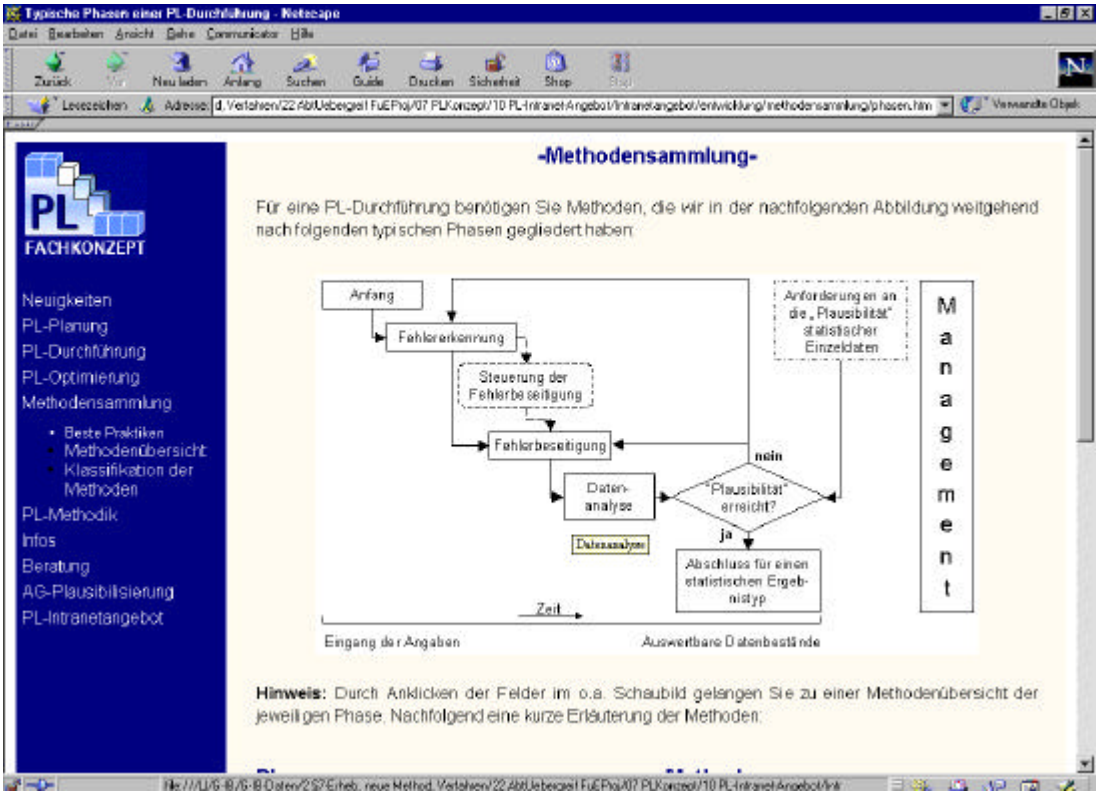

Figure 2: Screenshot of the start page of the method collection within the DE intranet pages showing the workflow of the execution of a data editing process. The boxes provide links to the methods associated with the selected work step.

IV. DE SCHEDULING: COLLECTION AND ASSESSMENT OF PLANNING CONDITIONS

7. The planning of an specific data editing process depends on a multitude of preconditions concerning the collection of data, the available human and production resources, the type and status of resulting data like micro data or aggregates and preliminary notices or final reports. In order to support the responsible department during the collection of these information and to make them available for the detailed planning of sub processes there was developed a prototype of an appropriate data base. Via explicit data base queries and XML interfaces these information can be imported by IT tools that utilise them during the realisation of subsequent sub processes (see next section for a description of the provided interfaces).

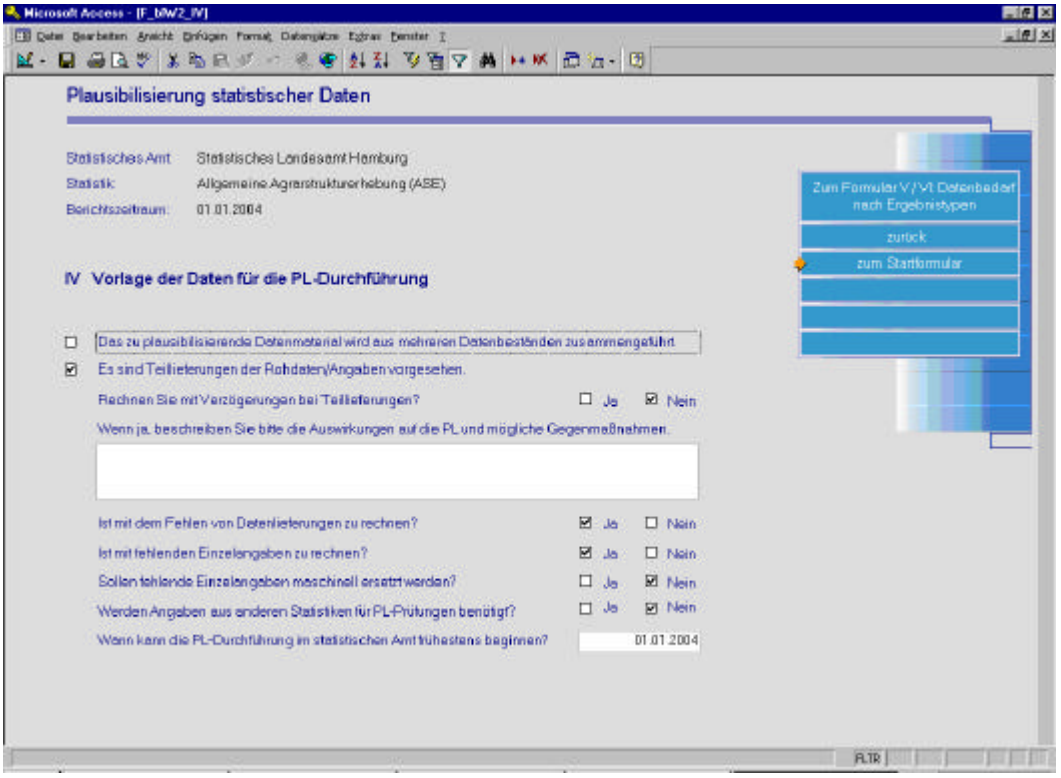

Figure 3: Screenshot of the DE scheduler showing a form for the specification of the status of the raw data before entering an editing process.

V. DE SPECIFICATION: THE EDITOR

8. The DE editor relates basic features concerning the specification of editing rules and data set descriptions with an object oriented data base containing among others the specifications of previously edited surveys and statistics. By this means specifications from completed data editing processes can be included in current application flows and thus support the reuse and harmonisation of specifications and increase the efficiency of the responsible department. Additionally the specifications can be made available throughout the complete process of data production resp. the particular IT tools.

9. Since the functionality of the DE editor is determined by the object-oriented structure of the underlying data base, figure 4 first of all reveals the aggregation of the involved objects and the cardinality of their (non total!) part-whole-relationships.

Within the program of official statistics every survey is associated with an one-to-one identifier. Following the aggregation arrows in reverse direction resp. de-referencing the corresponding inclusions, this key attribute allows for the access to the instances of the subsequent objects. Apart from data base queries concerning the immediate reuse of stored specifications during a DE editor session, this access is permitted to a wide variety of IT tools like for the generation of electronic questionnaires (incl. reuse and adaptation of variable specifications, derived plausibility checks and format definitions) or tabulation programmes using format specifications. Thus the DE editor may be considered as an user interface for the convenient enter of editing rules and data set descriptions as well as a front end of the underlying data base.

Basically a survey is considered as consisting of a set of topics and a corresponding data set description. Topics are disjoint sets of variables sharing a context of contents (like headers of an item block in a questionnaire). The union of topics to a superior topic is explicitly permitted and provides a useful

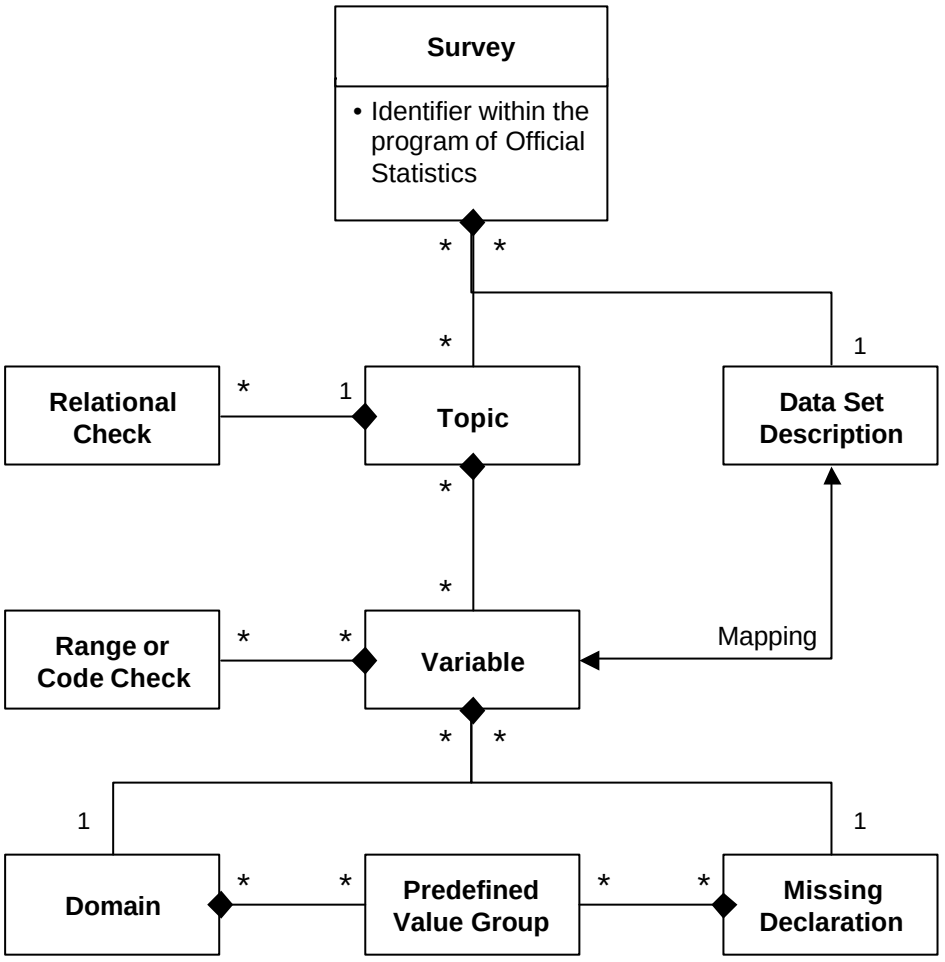

Figure 4: Aggregation of the objects provided by the DE editor (incl. cardinalities of the non total part-whole relationships, where "" denotes an arbitrary natural or null).*

method to structure the variables. These contexts induce relational data checks between the subsumed variables that are assigned to the topics. Analogous range and coding checks are related to the relevant variable. Since the reuse of specifications normally refers to both kinds of check procedures, topics are the first objects to be de-referenced.

The bi-directional arrow connecting the two objects "Variable" and "Data Set Description" denotes, that the data set description of a specific survey can be derived from the variable declarations and vice versa. This relationship is particularly helpful while reusing data set descriptions stored in the data base. To be specific figure 5 shows an register card of the DE editor providing variable definitions with variable names, data type declaration, field widths and write formats which determine the main features of the corresponding data set description. Combining these information with the mapping of the variable names onto the corresponding field names yields the constituents of a data set description (field names are entered during the specification of the first context the variable is assigned to). The other way round these fields may be initialised by a data set description either from a previous survey stored in the data base or delivered with the raw data.

10. Apart from structured data base access by common query languages the DE editor provides XML interfaces formatting the direct output streams. Since some IT tools and in particular analysis software like SAS or SPSS require specific input formats that cannot (and due to generality should not) be covered by the output declaration of the DE editor, there is a general I/O stream defined that provides the common

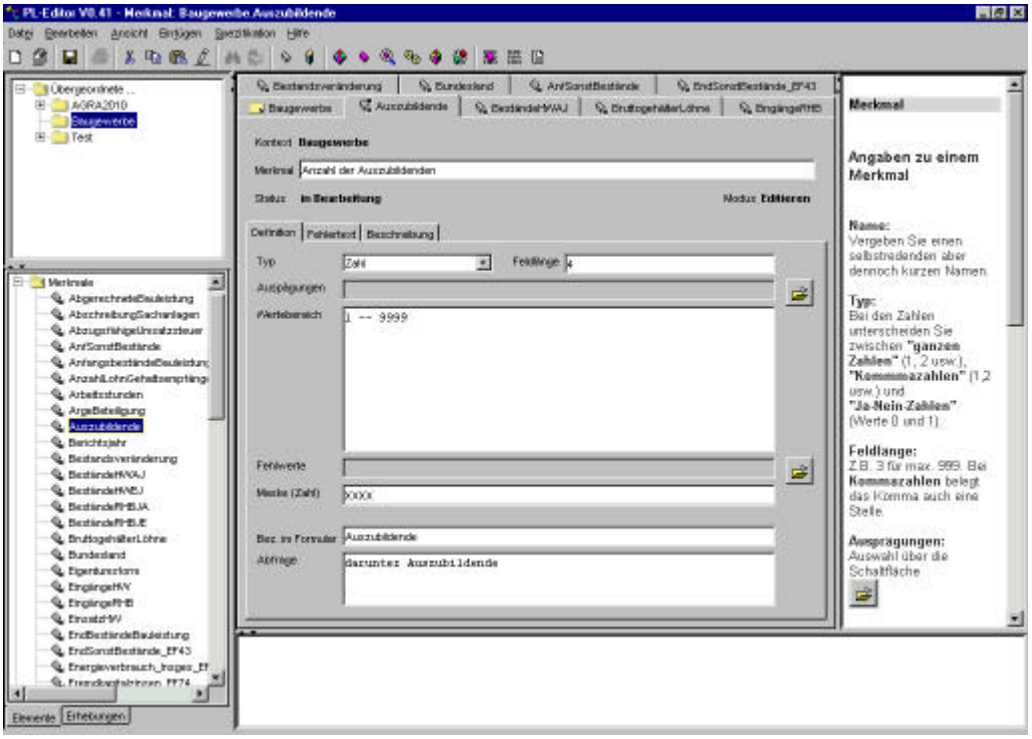

Figure 5: Screenshot of the DE editor showing a register card of the variable "Number of Trainees" providing specifications for the data type, the field width, the domain and the missing declarations of the variable. The two boxes on the left allow for the navigation between surveys (upper box) and within the selected survey. The box on the right shows the online help for the current work step (specifying a variable). The footer contains a box with log information about the execution of functions.

XML format. Via this gateway the DE editor also imports meta information about the data editing process from other IT tools like the DE scheduler in the previous section. The parameters and arguments required from the DE editor by the subordinated software can be more or less easily read by predefined or taylor-made XML parsers. The output of the parsers is then transcribed into executable source-code by automatic code generators. During the first phase of implementation only Java generators were introduced. Currently the interface between the DE editor and the software tools performing the error detection algorithms is defined.

11. In order to meet the "once-in-a-blue-moon" problem mentioned in the third paragraph, the DE editor provides a variety of help and documentation facilities. Of particular importance is the context sensitive online help in the right third of figure 5. It provides suggestions for work flows that can be seized by hyperlinks calling the demanded views and functions. Within a specified view by pushing the F1-key a context sensitive help function is available that provides information about currently activated input boxes. In addition the online help provides links to the intranet pages describing the method or work step in question.

VI. OUTLOOK AND FURTHER DEVELOPMENTS

12. As already mentioned in the introduction, the current implementations refer to sub processes concerned with the planning of data editing. During the subsequent phases the emphasis will be on the implementation of control methods as well as on methods of automatic error detection and correction. For

instance a selective data editing method is already implemented and tested in a SAS based simulation study and will be realised by a stand-alone IT tool within the next month.

The adaptation of existing and the development of new methods in data correction is focussed on the integration of case based editing rules and distribution based imputation algorithms. Centre of reference is to realise the theory of multiple imputation under the constraints of practical data editing in official statistics. In doing so a two step strategy is accomplished: In the medium term the integration is realised in a pilot project in the annual cost structure survey in the construction industry, where existing software for the generation of multiple imputation (IVEware by Raghunathan) will be introduced for the first time in the Federal Statistical Office. In the long run it is planned to replace these tools by tailor made solutions extended with non linear approximation methods like multilayer perceptrons. First simulations of multiple imputations with multilayer perceptrons are already carried out on incomplete observations drawn from continuous random distributions and showed promising results.