STATISTICAL COMMISSION and Working Paper No. 3
ECONOMIC COMMISSION FOR EUROPE English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic I: Management and evaluation of editing and imputation procedures

THE PLANNING OF DATA EDITING

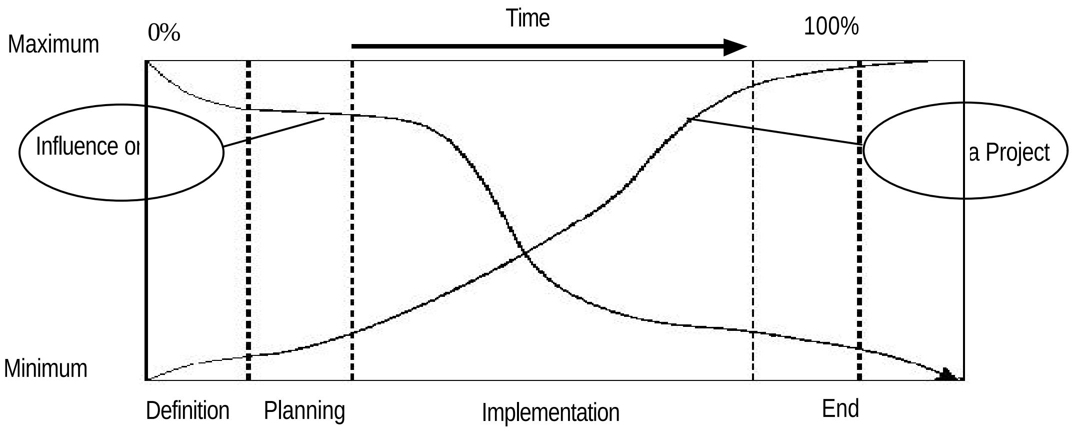
Submitted by the Federal Statistical Office, Germany[1]

Invited paper

I. INTRODUCTION

1. Many National Statistical Institutes (NSIs) today face two main contrary challenges like increasing user demands for statistical data and (continuous) budget cuts. They represent a permanent incentive for the improvement of survey activities. Efficiency of survey activities leads to lower costs which can be influenced by planning activities as shown in figure 1:

Figure 1: (Influence on) Costs during the Run of a Project [1]



2. Due to the contrary developments mentioned in the first paragraph the number of conditions which have to be borne in mind during the planning of data editing increases. From this development the danger may arise that data editing may be rendered inefficient because of poor planning. Planning of data editing is heavily influenced by survey specific aspects but in addition to those general methods and procedures should be implemented to achieve more efficient survey operations and a standardization of planning activities with the benefit of a better comparability of surveys. Project management may offer useful methods for the planning of data editing.

3. The aim of this contribution is therefore to provide:
- an overview of the planning of data editing with adapted project management techniques and
- some reflections on selected planning activities.

[1] Prepared by Elmar Wein.

4. The contents of the following sections are influenced by the discussion of a German task force which is developing guidelines for data editing.

II. OVERVIEW OF THE PLANNING OF DATA EDITING

II.1 Assumptions made for the planning of data editing

5. Users' demands for data quality influence data editing. There is no uniform definition of data quality, but with regard to the planning of data editing the focus may be put on the commonly used criteria "timeliness", "accuracy" and "clarity, accessibility".[2] Clarity is an important criterion for the planning of data editing because data editing must in many cases provide information for informative and user friendly quality indicators.

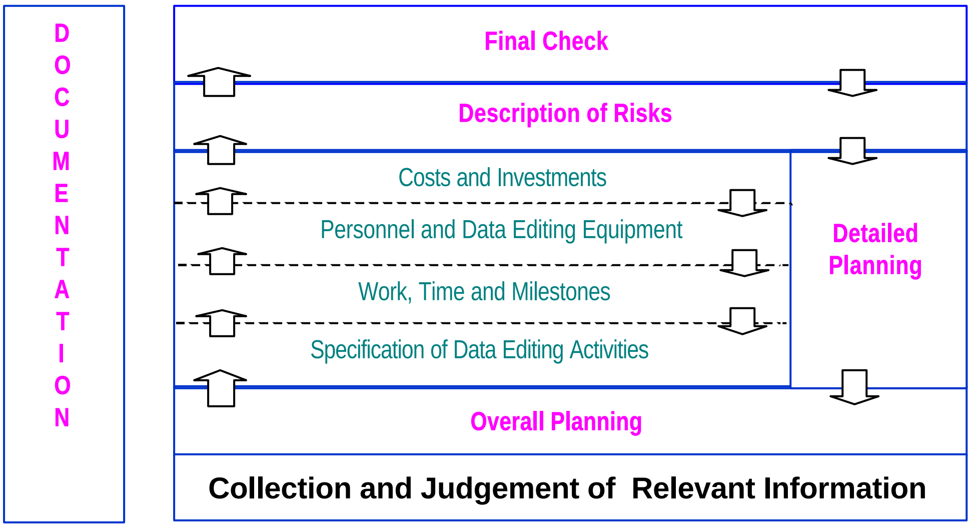
6. Current developments in the theory of organization promote the reengineering of business activities by the implementation of process organization and management.[3] For that reason data editing activities are combined in data editing processes.[4] A data editing process contains the imputation or modification of statistical data as a result of logically connected activities. The design of a data editing process reflects the individual view of an organizer. Data editing processes are designed in such a way as to contribute to the dissemination of statistical results by short runtimes, low consumption of resources and userfriendly documentation. They possess the absolutely necessary interfaces to other survey processes and complex data editing processes can be divided into logically separated sub-processes. A process owner is defined for every data editing process who needs information, methodological and subject matter knowledge and adequate equipment.

7. The execution of a survey can be described by seven primary processes: "clarification of the requirements of statistical results", "survey preparation", "data collection", "data processing", "data analysis", "dissemination" and "survey optimization". Data editing consists of three processes: "preparation of data editing", "execution of data editing" and "optimization of data editing", which are sub-processes of various primary processes. The planning of data editing as a sub-process of the preparation of data editing happens during the primary process "survey preparation". It needs information coming from other processes like "clarification of the requirements of statistical results", questionnaire design and methodological knowledge which is available in data editing guidelines. Information created by the planning of data editing, i.e. specification of edits is needed for electronic data processing (EDP). Documentation of the plans and specifications are needed for optimization of data editing which may also influence methodological guidelines for data editing.

II.2 Planning sub-processes

8. The planning of data editing can begin when the design of the questionnaire is finished and record / file descriptions are available. Ideally the staff who are responsible for the questionnaire design also plan the data editing. The planning of data editing consists of various stages as illustrated in figure 2:

Figure 2: The Flow of the Planning of Data Editing


DOCUMENTATION
Final Check
Description of Risks
Costs and Investments
Personnel and Data Editing Equipment
Work, Time and Milestones
Specification of Data Editing Activities
Detailed Planning
Overall Planning
Collection and Judgement of Relevant Information

9. Planning of data editing starts with the collection and judgement of relevant information.[5] Subject matter specialists should gain knowledge about the main tasks and relevant conditions of the data editing to be planned. Beyond that, an adequate documentation should make inconsistencies in judgements obvious and enable colleagues to bring in their experience.

10. On the basis of the evaluated information a consistent data editing strategy is developed by a top-down approach within the overall planning. It divides data editing into sub-processes, sets preconditions for instance concerning data editing methods, process durations and costs and is documented in a structural plan.

11. The ensuing detailed planning is characterized by changes between planning activities and reviews to promote internal monitoring. Discrepancies between the detailed plans must be solved by the examination of all preceding plans which may finally also influence the data editing strategy. The detailed planning consists of the specification of data editing activities like edits, data evaluation, electronic data processing and documentation corresponding to the overall planning and is performed in a bottom-up approach. On the basis of specified data editing activities the detailed planning of work, time, personnel, equipment, costs and investments follows. These procedures are terminated when a balance is achieved between the "requirements of statistical results", "resources", "survey organization", "available time", "data editing effort" and if there are no inconsistencies between the different detailed plans. After the detailed planning the description of risks is necessary to highlight the conditions which may cause a failure of the planned conduct of data editing. During the final check as a last step inconsistencies within and between the different detailed plans should be detected. Important documents of the detailed planning are the specification of data editing activities, work manuals, the time and milestone table, cost and investment plans and the description of risks.

12. It is assumed that all documents will be held together with survey specific metadata like survey contents and specifications of statistical results in data bases. These systems are accessible to all members of a survey managing unit so that they can fill them when data will be created. The data bases enable a multipurpose use of metadata such as the specification of edits on the basis of specified characteristics and their analysis for instance.

III. ASPECTS OF PLANNING ACTIVITIES

13. The previous paragraphs have shown the complexity of the planning of data editing, so it is not possible to describe all sub-processes in detail in this contribution. Some stages seem to be more relevant

for a successful planning: e.g. the collection and judgement of relevant information, the overall planning and the judgement of the specified edits.

III.1 Collection and judgement of relevant information

14. The collection and judgement of relevant information determine the knowledge of the staff about the data editing to be planned and describe interfaces to other survey processes. A crucial procedure of this sub-process is the transformation of the demand for statistical results into operative preconditions for data editing. It is assumed that a NSI possesses a dissemination strategy which defines certain categories of statistical results, e.g. "Tendencies", "Preliminary Results", "Standard Results" and "Microdata Files". These categories mainly differ concerning their timeliness, accuracy and statistical data to be edited. Tendencies incorporate a high degree of timeliness and a low degree of accuracy and microdata files vice versa. A subject matter unit proposes in accordance with the dissemination strategy and legal duties timeliness for the types of statistical results e.g. in days, weeks and indicators for accuracy e.g. variation coefficients, relative standard errors or relative changes of the respective statistical results.[2] These descriptions are offered to users during the survey process "clarification of the requirements of statistical results" and for instance adapted at user conferences during the process "survey optimization".

15. With the knowledge of users' demands, survey contents and estimations of response behaviour the survey management should provide the following list which forms the external aims of data editing:

Figure 3: Extract of a List with External Aims of Data Editing - Example

Type of Statistical Result (r)	**Characteristics** **(c)**	**Termination of a sub-process of data editing when ...**	**Duration** **[months], (d_{cr})**	**Additional Notes**
1	2	3	4	5
Tendency	*1. Number of Persons in a Household* *2. ...*	*1. Relative Change of the Corresponding Statistical Result ≤ 20 %* *2. ...*	*1. 4* *2. ...*	

16. The information in column 1 determines the data editing strategy. The information of column 4 is used as a precondition for the detailed planning and must be confirmed by the time scheduling. The information of column 3 is used as an indicator which possesses a strong relation to the accuracy of a statistical result. With the information of column 2 and the revised one of column 4 the priority of a characteristic Pt_c can be computed to be used for priority setting of data editing:

(3.1) $$Pt_c = \max_{r=f(c)} \left(\frac{d_{max}}{d_{cr}} \right), d_{cr} > 0 \text{ and } d_{max} = \max_r (d_{cr})$$

17. Collection and judgement of relevant information concerns subject matter, EDP and organizational aspects. Subject matter aspects can be taken from legal acts or treaties with clients; in case of primary statistics from survey instruments and other survey documents, e.g. interviewer manuals; in the case of secondary statistics the usage of registers and processed data play a dominant role for the collection of information. That information should provide an impression of the data editing scope, necessary qualification of the personnel and information needed for the specification of edits.

[2] The European Commission Regulation No 161/1999 contains an example of a legal duty for the evaluation of statistical data.

18. EDP aspects concern the available hard- and software and the requirements set by the survey. Information about the existing personnel and their qualification, the finances and delivery times of data (in a decentralized statistical system) belong to organizational aspects.

19. Important sources of information about similar or equal edits may be gained from pretests, pilot studies and similar or preceding surveys. As pretests are used for the testing of questionnaires in samples with mostly "critical" respondents they only provide information about possible errors.

20. Pilot studies are a good opportunity for tests of data editing because they may deliver useful information on error frequencies, data editing effort and perhaps on efficiency of edits. The report of a pilot study should therefore contain information about:

- Error descriptions with involved characteristics,
- Error frequencies,
- Effort needed for the correction of errors,
- Recommendations concerning tested edits,
- Information about the usefulness of error descriptions, instructions for corrections and work manuals.

21. Similar or preceding surveys may deliver information for identical characteristics and they help experienced staff to estimate the effort for data editing. Information about error frequencies, the effort needed for correction and specification of edits can be used for data editing in the case of identical characteristics.

22. After the collection of information statisticians have to sort positive and negative conditions according to their relevance for data editing. In many cases the demand for statistical results cannot be fulfilled with the given resources or the influence of negative preconditions must be restricted. For those reasons statisticians should set priorities or focuses among conditions which can be influenced such as survey organisation, available resources, planned effort for data editing, available time and planned statistical results.

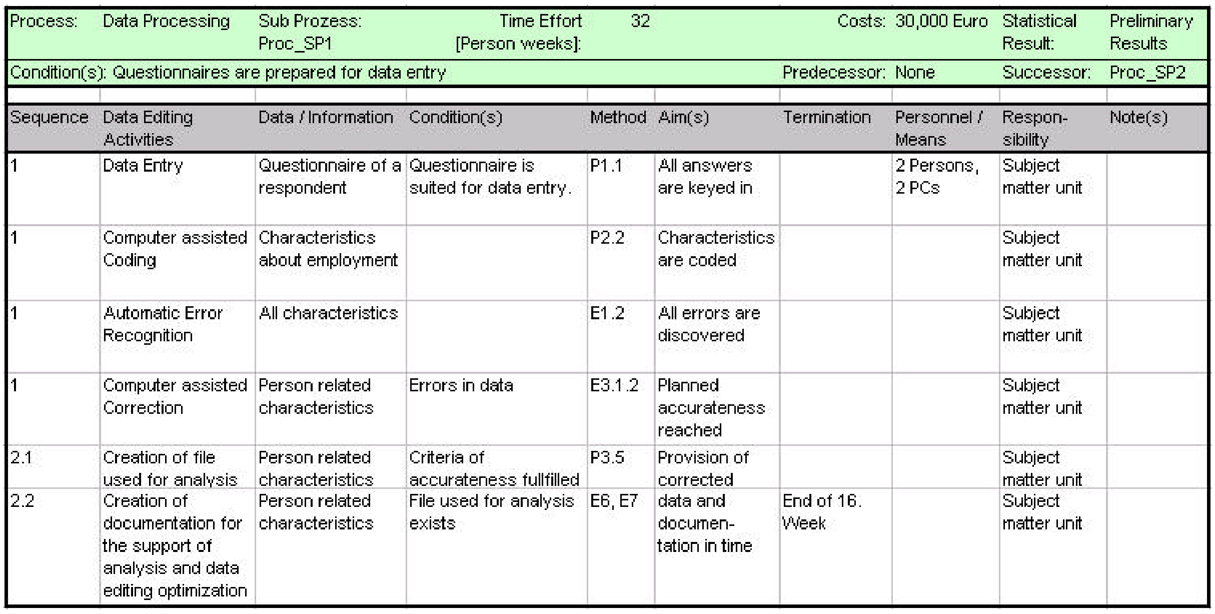
23. After those steps statisticians may have gowned a sound impression of the data editing to be planned so that they might be able to name possible risks. In case of improvement of existing surveys internal aims such as the reduction of the data editing effort or increasing efficiency of internal process organization may be formulated at the end of this procedure.

III.2 Overall planning – development of a data editing strategy

24. The development of an output-oriented, consistent data editing strategy for the processes "data collection" and "data processing" can start when the collection and judgement of relevant information has been terminated. This process needs especially information as shown in figure 3, information about cost limits, available financial means and data editing guidelines. In addition to structural plans preliminary time tables with durations and deadlines for sub-processes are set. They represent preconditions for the ensuing detailed planning and must be adapted if the detailed planning results cannot be kept.

25. Data editing activities during the process "data collection" are in general divided into processes performed by respondents, interviewers and those performed in NSIs. Data editing operations during the process "data processing" may be divided in sub-processes in accordance with the needed types of statistical results. Figure 4 contains an extract of a structural plan for the conduct of data editing during the process "data processing".

Figure 4: Extract of a Structural Plan of Data Editing - Example



Process:	Data Processing	Sub Prozess: Proc_SP1	Time Effort [Person weeks]:	32		Costs:	30,000 Euro	Statistical Result:	Preliminary Results
Condition(s): Questionnaires are prepared for data entry						Predecessor:	None	Successor:	Proc_SP2
Sequence	Data Editing Activities	Data / Information	Condition(s)	Method	Aim(s)	Termination	Personnel / Means	Respon-sibility	Note(s)
1	Data Entry	Questionnaire of a respondent	Questionnaire is suited for data entry.	P1.1	All answers are keyed in		2 Persons, 2 PCs	Subject matter unit	
1	Computer assisted Coding	Characteristics about employment		P2.2	Characteristics are coded			Subject matter unit	
1	Automatic Error Recognition	All characteristics		E1.2	All errors are discovered			Subject matter unit	
1	Computer assisted Correction	Person related characteristics	Errors in data	E3.1.2	Planned accurateness reached			Subject matter unit	
2.1	Creation of file used for analysis	Person related characteristics	Criteria of accurateness fullfilled	P3.5	Provision of corrected			Subject matter unit	
2.2	Creation of documentation for the support of analysis and data editing optimization	Person related characteristics	File used for analysis exists	E6, E7	data and documen-tation in time	End of 16. Week		Subject matter unit	

The estimated time effort of the process "Proc_SP1" amounts to 32 person weeks of a subject matter unit and a CADI application with immediate computer assisted error correction is used for data editing (sequence no. 1). When a predefined accuracy is reached the sub-process ends – finally in the 16th week with the creation of a data file and corresponding documentation needed for data analysis and optimization of data editing. After "Proc_SP1" the sub-process Proc_SP2 fellows - for example data editing for standard statistical results. The column "Method" contains shorthand expressions which represent categories of methods used for the conduct of data editing processes.[6]

III.3 Specification of data editing activities - judgement of specified edits

26. As the planning of data editing consists of a lot of different activities the judgement of the detailed plans will be facilitated by benchmarks, which should enable comparisons between "similar" surveys. As not all effects are explicable by benchmarks rather homogenous groups of surveys should be created. They may be classified by periodicity, types of data collection and the attribute characteristic relation AC which defines the relation between the sum of all attributes Y_s and the sum of all characteristics X_s:[3]

$$AC = \begin{cases} \dfrac{Y_s}{X_s} & ; X_s > 0 \\ 0 & ; else \end{cases} \qquad (3.2)$$

Codes of a classification are treated as dependent characteristics which represent the digits. A classification may consist of 7 main groups on the first level which can be treated as one characteristic with seven attributes. There are 5 groups on the second level per main group so that the second digit of a code can be treated as another characteristic with 5 attributes.

27. Besides this classification a lot of differences due to the size of survey contents and planned statistical results may still occur. Benchmarks should restrict the effects of those differences. As the planning of data editing fixes edits, work packages, time effort and costs benchmarks should provide information about all those aspects and make relations between them obvious.

28. Benchmarks for the judgement of specified edits may form the basis of a consistent system of benchmarks for the planning of data editing. They should make differences concerning edits of various surveys apparent and facilitate comparison. A top indicator [7] for the analysis of edits may be the

[3] In case of constant variables it is assumed that respondents generally give one item of information.

specification scope SC which is the ratio between the sum of all edits Z_s and the sum of all characteristics X_s – including needed characteristics from other surveys and computed characteristics, excluding information given on open-ended questions:

$$(3.3) \qquad SC = \begin{cases} \dfrac{Z_s}{X_s} & ; X_s > 0 \\ 0 & ; else \end{cases}$$

Instead of the sum of all characteristics the sum of all attributes may be used to raise the information of this benchmark.

29. The specification scope should be completed by the share of manual edits ME which is defined as the ratio between manual edits Z_m and the sum of all edits Z_s:

$$(3.4) \qquad ME = \begin{cases} \dfrac{Z_m}{Z_s} & ; Z_s > 0 \\ 0 & ; else \end{cases}$$

To achieve better information the specification scope and all succeeding benchmarks may be computed separately for manual edits. Another aspect may be a distinction between characteristics of one survey and those of other surveys to gain an overview of the dependence from other surveys.

30. Different specification scopes should be explained by further benchmarks. Surveys normally contain different measurement levels u which may influence the number of edits. Measurement levels may be defined with $u = 1$ for nominal characteristics, $u = 2$ for ordinal, $u = 3$ for quantitative and $u = 4$ for open characteristics. The structure plausibility SP_{1+2} is the ratio between the number of coding edits Z_d and the sum of nominal X_1 and ordinal characteristics X_2:

$$(3.5) \qquad SP_{1+2} = \begin{cases} \dfrac{Z_d}{X_1 + X_2} & ; \quad X_1 + X_2 > 0 \\ 0 & ; \quad else \end{cases}$$

SP_{1+2} may not be greater than 1. With the number of range edits Z_W and the number of quantitative characteristics X_3 the structure plausibility for quantitative characteristics SP_3 will be defined similarly to formula (3.5). The maximum of it is 1 or 2 – depending on the way how lower and upper bounds for range edits are specified.

31. For the comparison of consistency edits between "similar" surveys the interplausibility IP defined as the ratio between the number of consistency edits Z_b and all edits Z_s may provide useful information:

$$(3.6) \qquad IP = \begin{cases} \dfrac{Z_b}{Z_s} & ; Z_s > 0 \\ 0 & ; else \end{cases}$$

32. Reasons for a high interplausibility may be a high number of involved characteristics, the multiple use of characteristics in different consistency edits and a high scope of combinations. The number of involved characteristics IC is defined as the ratio between the number of characteristics which are involved in consistency edits X_b and the sum of all characteristics X_s:

$$(3.7) \qquad IC = \begin{cases} \dfrac{X_b}{X_s} & ; X_s > 0 \\ 0 & ; else \end{cases}$$

33. The multiple use of characteristics in consistency edits MC measures how often a characteristic c is used in different consistency edits b. With the frequency of a characteristic in a consistency edit $f_{c,b}$ it may be defined as:

$$(3.8) \qquad MC = \begin{cases} \frac{1}{Z_b} \cdot \sum_{c=1}^{Z_b} \sum_{b=1}^{B} f_{c,b} & ; Z_b > 0 \\ 0 & ; else \end{cases}$$

34. A high scope of combination SC may be a third reason for a high interplausibility. It may be defined by the number of characteristics i_b which are involved in a consistency edit b:

$$(3.9) \qquad SC = \begin{cases} \frac{1}{Z_B} \cdot \sum_{b=1}^{B} i_b & ; Z_B > 0 \\ 0 & ; else \end{cases}$$

References

[1] Jacques Boy, Christian Dudek, Sabine Kuschel, (1996). "Projektmanagement". Offenbach, 37.

[2] Yves Franchet (1998). "Verbesserung der Qualität des ESS". DGINS-Konferenz, Stockholm, 18. Bernard Grais (1998). "The Future of European Social Statistics". Mondorf Seminar, pp. 28-31. Martin Collins, Wendy Sykes (1999). “Extending the Definition of Survey Quality”. Journal of Official Statistics, Vol. 15, No. 1, pp. 57-66.

[3] Bernd W. Wirtz (1996). "Business Process Reengineering – Erfolgsdeterminanten, Probleme und Auswirkungen eines neuen Reorganisationsansatzes". Zeitschrift für betriebswirtschaftliche Forschung, 48, pp. 1023 – 1037.

[4] Manfred Schulte-Zurhausen (1995). "Organisation". München, 41pp. Günter Schmidt (1999). "Methoden des Prozess-Managements". WiSt, 9, pp. 241-245. Verein Deutscher Ingenieure, Deutsche Gesellschaft für Qualität (1998). "Total Quality Management Prozesse". VDI/DGQ 5505 (Entwurf), Düsseldorf, pp. 2-17.

[5] Georg A. Winkelhofer (1997). "Methoden für Projektmanagement und Projekte". Berlin, pp. 121-215. Heinrich Keßler, Georg Winkelhofer (1999). "Projektmanagement". Berlin, pp. 162-180.

[6] Dania P. Fergusson (1994). "An Introduction to the Data Editing Process". Statistical Data Editing, Volume No. 1, Methods and Techniques, pp. 1-9.

[7] Peter v.d. Lippe (1993). "Deskriptive Statistik". Stuttgart, pp. 20-23. Zentralverband Elektrotechnik- und Elektronikindustrie e.V. (1989). "ZVEI-Kennzahlensystem". Frankfurt/Main, pp. 27.