New approaches to data processing integration in North Rhine-Westphalian Statistics based on BLAISE III

Markus Broose, Frank Merks, Thomas Pricking, data processing and statistical office of the State of North Rhine Westphalia - Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen - (LDS NRW), Germany

1. Tasks and structure of the statistical network in the official statistics of the Federal Republic of Germany

Official statistics in the Federal Republic of Germany are supplied by the federal and state statistical offices, which compile regional, supra-regional and national data. All European, federal and state statistics are compiled in accordance with the prescriptions of European, national or state law. Federal statistics comprize an annual approx. 180 sets of statistics, supplemented in North Rhine Westphalia by a further 30 so-called "coordinated sets of state statistics". These are compiled by all states of the Federal Republic with the federal statistical office usually exercising a coordinating function, as comparability of results is in both the federal and state interests. In addition to this, specific state surveys are carried out in the various states: in North Rhine Westphalia these amount to approx. 30 sets of statistics.

Federal statistics are compiled in a division of labour between the 16 state statistical offices and the federal statistical office. In close consultation with the state statistical offices the federal statistical office undertakes the methodological and technical preparations, coordination of the various sets of statistics and compilation and publication of the end results. The state statistical offices undertake the acquisition and processing of the statistical data together with evaluation and distribution of the results relating to their own particular state. Exceptions to this rule are a few sets of statistics which are dealt with centrally by the federal statistical office. The state statistical offices act in the capacity of independent state authorities which are not subject to control by the federal statistical office. They are funded by the states themselves and accordingly bear a large part of the financial burden of compiling federal and European statistics.

Given this structure, both federal statistics and the coordinated state statistics must be compiled in a uniform manner by all concerned so as to assure regionally and technically comparable results. That means that the

same methodological principles and procedures must be employed in all the statistical offices. In order to meet this requirement, the federal and state statistical offices have formed a "network" to develop and implement the necessary measures in close cooperation with one another. This statistical network relates not only to methodological and technical issues but also to collaboration in the creation of IT applications for processing the data acquired in surveys, ie. so-called "network programming".

Network programming aims at the achievement of two goals :

- *Standardization of methods wherever and in whatever division of labour the statistics are compiled.*

- *Optimum cost effectiveness through division of labour.*

In order to attain these goals, the statistical network has laid down criteria for the organization and programming of IT applications which must be adhered to by all parties involved :

- *The technical rules are laid down by the relevant technical committees.*

- *Specification guidelines for validation and tables in the form of a semi-formal language and for the technical/organizational implementation of the rules define a common basis for procedure.*

- *The employment of IT is controlled and coordinated by the information technology working party AKIT. All the statistical offices are represented in this working party.*

- *Standardized working procedures are employed in all statistical offices.*

- *At each stage of the work process the same program is used everywhere.*

- *The programs must be easily portable, as a variety of different platforms are in use within the network.*

- *As a rule, the requisite software is created, documented and maintained by one statistical office. All the other offices then receive the applications free of charge and load them into their systems. Platform-specific functions (job control, printer drivers etc.) must be adapted by the user.*

The federal and state statistical offices together allow an annual capacity of 80 man years for network programming. The synergy effects achievable by means of this division of labour are enormous. If the network did not exist federal statistics in their present form, ie. produced in an organized division of labour between the statistical offices, would scarcely be feasible.

2. Technical basis and perspectives of network programming

Network programming has been in existence since the beginning of the 1960s. Accordingly, mainframe computers (IBM/MVS and SNI/BS2000) have constituted the backbone of the system for more than three decades. For reasons of portability and efficiency the statistical offices have in the past used assemblers and assembler macros as programming tools. Such applications, which are still extensively used today, are run exclusively as batch jobs. In recent years many assembler programs have been successfully replaced by a 4GL language for evaluation and tabulation of results (SPLV, developed by the statstical network) which is tailored to the technical jargon of statisticians. This software is also job driven.

Whereas, in the past, batch oriented procedures were largely equal to the task of effectively supporting the processing of statistics, today's demands can only be met using modern methods. Statisticians' most urgent demands are for interactive solutions, more versatile and more decentralized technology together with faster and better presentation of results. Given increasingly stringent public sector budgets, cost cutting has clearly also become an important factor in the field of statistics.

Germany's public sector statisticians are making efforts to achieve these goals in two ways :

- *The ADABAS data base system using the programming language NATURAL has been introduced as a development tool for interactive mode programming. ADABAS is primarily a mainframe system but can also be run on Unix and PCs. ADABAS was quick to prove its merit in register applications for the production industry, trade and agriculture and also for statistical information data bases. In addition to this, ADABAS applications are used to run statistics processing including data validation in interactive mode.*

- *Micro-computers with their extensive range of capabilities are to constitute the other main pillar of our drive to modernize computerized statistics processing. Thanks to local area networks PCs can nowadays communicate with one another and, if necessary, be quickly connected to a mainframe, ensuring efficient data transfer at all times. The software industry has a wide range of development tools and ready-to-use applications on offer which are of great service in the preparation and processing of statistics. However, these programs all have the disadvantage that they are not, or are only to a limited extent, designed to cope with the special needs of official statistics. For example, they lack tools for error handling, coding of plain text and job sequencing. With regard to these points in particular, CBS Netherlands' BLAISE survey processing system is far ahead of the field and accordingly became standard in network programming in 1991, since when it has been used in a variety of surveys.*

There is now the evidence of a number of surveys to show that switching from batch oriented procedures to interactive mode solutions has substantial advantages.

Owing to organizational considerations beyond the scope of the present article, the 1994 micro-census, the most important sample survey for

population and social statistics[1], was carried out in North Rhine-Westphalia using approx. 45% conventional methods, ie. manual handling of paperwork and data registration and validation in batch runs on a host computer, and 55% using a BLAISE interactive program which already contained most of the validation checks. The "raw data" gained by both methods was subjected to final validation using batch programs on the host computer of the statistical network. This showed that only half the number of runs was necessary for the BLAISE material and that the fraction of implausible cases in the first validation run was only 3% of the error quota ascertained for the data produced by conventional means.

Running under ADABAS the metadata driven application DAMAST, used for processing building trade statistics, also demonstrated the usefulness of interactive solutions.

Similar results have also been obtained by other statistical offices. Nonetheless, it is not at present anticipated that all statistical offices will be switching quickly and completely to processing all their statistics in interactive mode. The conventional procedures will continue to be used in the statistical network for a certain time to come. This begs the question as to why this should be so, despite the advantages of interactive solutions and the favourable response of users to them. The reasons are complex and closely connected with the autonomy of the 17 statistical offices making up the statistical network.

- *As a rule, modernization projects require high initial capital expenditure and this poses a problem in times of budget cutbacks. The budgets of the statistical offices come out of federal or state funds and there is strong pressure to economize. Accordingly, investment in the new data terminals required for the implementation of interactive solutions has to be medium-term.*

- *As yet there are not enough programmers trained in the new techniques available. Resources for software development are accordingly limited.*

- *The planning of modernization projects in the statistical network presupposes agreement in detail between all parties involved. Coordination of the various specialist departments and IT staff easily results in project timescales of several years, during which statistics will necessarily continue to be produced by conventional methods.*

- *A massive introduction of interactive techniques will fundamentally change the statisticians' workplace. Work flows, organizational structures and areas of responsibility will all have to be redefined. The effects on personnel, ranging from reassignment and transfer to redundancy, may well be considerable. On the other hand, no statistical office will voluntarily accept losses of qualified staff, particularly as the demands made on statistics are becoming increasingly heavy.*

[1] *Micro-census is a 1% area random sampling procedure. In North Rhine-Westphalia in 1994 76.000 households were polled.*

Modernization measures must accordingly be a matter of long-term planning.

3. BLAISE in the data processing and statistical office of the State of North Rhine-Westphalia (LDS NRW)

Since the beginning of the 1990s LDS NRW has been making strenuous efforts to improve the compilation of statistics. At first concerned chiefly with better and faster tabulation of corrected data (statistical problem solving procedure), the measures envisaged by LDS are now aimed at the changeover to interactive procedures. As a member of the statistical network LDS must continue to participate in common procedures and is unable to "go it alone". On the other hand, LDS is reluctant to miss the opportunities offered by modern technology. Accordingly, a way out of the dilemma has had to be found.

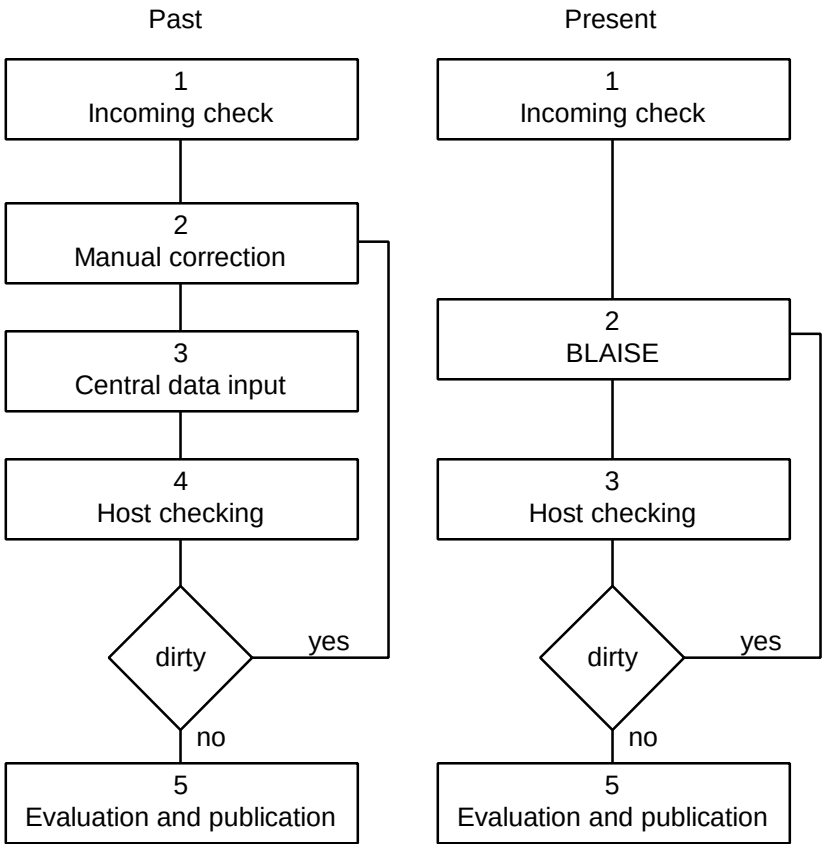
For one thing, LDS urges the increased use of interactive solutions in the statistical network. An example of this is the development of the ADABAS/NATURAL procedure for various surveys coordinated by LDS and its implementation within the statistical network. Whenever completely new procedures are developed for the statistical network consideration is to given to the question of whether PC solutions are technically and economically feasible. In the relevant committees LDS supports programming with BLAISE.

For another, it is urged that a particular way of using BLAISE be adopted which combines PC and conventional procedures. This intermeshing of PC and host computer applications will be referred to henceforth as "data processing integration". The following kinds of statistics are especially well suited to this procedure :

- *Statistics which have hitherto been validated, coded and registered using conventional methods.*
- *Statistics that require extensive coding.*
- *Statistics which, in accordance with statistical network planning, are processed manually up until batch validation.*
- *Statistics for which no changeover to interactive procedures is planned by the statistical network.*

The task of the BLAISE program will be to effect validation and coding during input itself with the result that, in theory at least, error-free material is available as soon as the last questionnaire has been entered. This material - where applicable, together with similar data from external sources (EDI, tapes, floppies) - will then be subjected to a final validation on the host computer using the standard procedures laid down by the statistical network. In this way the material produced with the aid of BLAISE will be integrated in the standardized process of the statistical network. This procedure is illustrated in the chart given in Fig. 1.

Fig. 1 : Chart showing data processing integration using BLAISE III



Two objects can be achieved :

- *Full exploitation of the technical advantages of BLAISE while preserving the standardized method of the statistical network.*
- *Faster and more cost efficient processing*

The 1995 survey of salary and wage structure will serve as an example of the integration process and its resultant benefits.

4. 1995 survey of salary and wage structure (SWS)

The survey of salary and wage structure provides information on employees and their earnings, the nature and amount of statutory deductions and training and qualification characteristics. The survey is carried out on a random sampling basis in firms of more than 10 employees in the production industry, trade, banking and insurance. In 1995 this involved approx. 4,000 firms in North Rhine-Westphalia. Three different lists were used in the survey:

- *the company data sheet*
- *a list for wage earners*
- *a list for salaried employees.*

Prior to the main survey a preliminary poll was carried out in the firms concerned. The preliminary poll provided information on, for example, the collective bargaining agreements applicable to the firms, the number of employees and the way in which the firms wished to provide the information (questionnaires or data media). This information was processed into coding lists by LDS NRW and the federal statistical office which then formed the basis for parts of the validation procedures employed in the main survey.

Every firm was required to fill in a company data sheet with no omissions. A number of lists distributed for wage earners and salaried employees varied with the total number of people employed by the firm: The bigger the firm the lower the proportion of employees selected. On the forms used for the lists data for a maximum of 10 employees could be entered, with the result that medium sized and, particularly, large firms needed considerably more than one form.

This necessitated the creation of a system of data modelling and data input control that was versatile enough to cope with the varying nature and scope of the lists. Table 1 and Fig. 2 provide an illustration of the structure and scope of data. The flow chart given in Fig. 3 illustrates the individual steps of the procedure.

Tab.1 : Technical description of the model HSWS1_0X

- *BBetriebssatz*
  - *BArbeiterbogen*
    - *BArbeitertabelle*
      - *BArbeitersatz*
  - *BAngestelltenbogen*
    - *BAngestelltentabelle*
      - *BAngestelltensatz*

Overall counts

Number of uniquely defined fields[2]	*109*
Number of elementary fields[3]	*102*
Number of defined data fields[4]	*4788*
Number of defined block fields[5]	*7*
Number of defined blocks	*7*
Number of embedded blocks	*0*
Number of lock instances	*282*
Number of key fields	*1*
Number of defined answer categories	*3*
Total length of string fields	*1964*
Total length of open fields	*0*
Total length of field texts	*2819*
Total length of value texts	*405*
Number of stored signals and checks	*6168*

Data fields	*Number*	*Length*
Integer	*3719*	*2010*
Real	*808*	*4428*

[2] *All the fields defined in the FIELDS section*

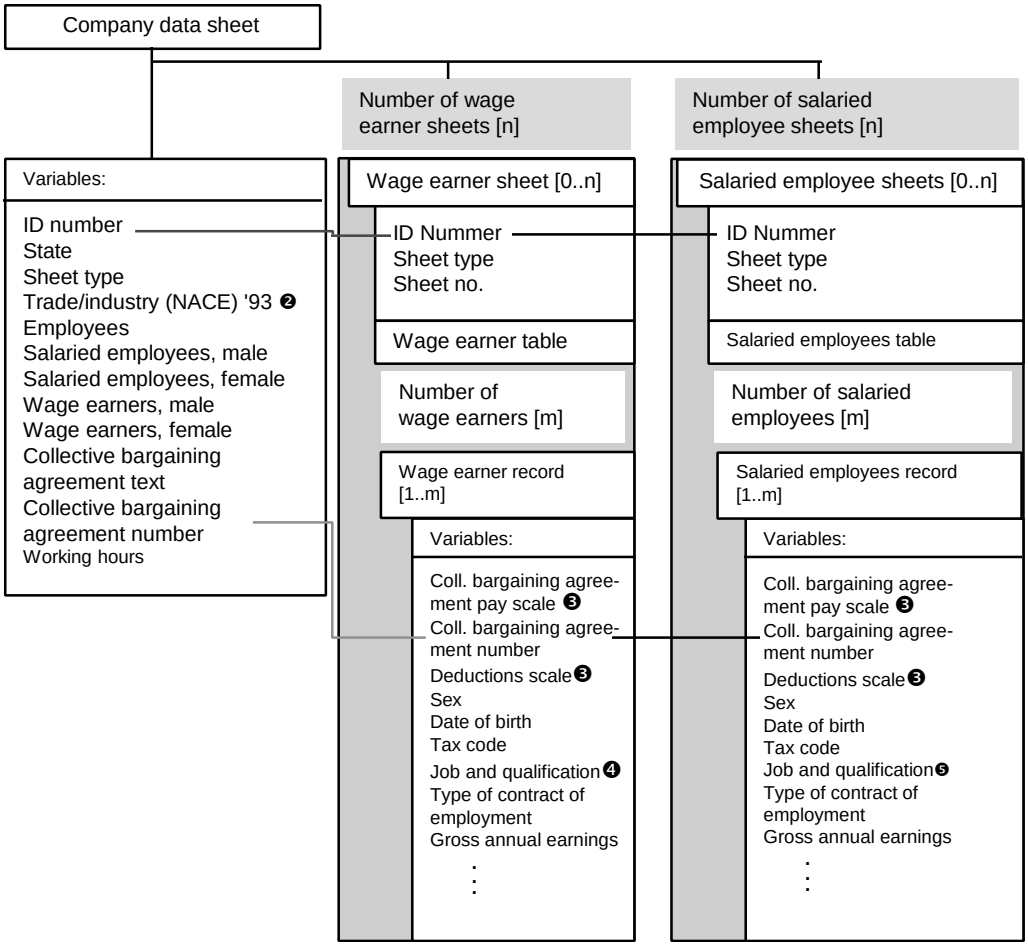
[3] *All the fields defined in the FIELDS section which are not of type BLOCK*

[4] *Number of fields in the data files (an array counts for more than one)*

[5] *Number of fields of type block*

Enumerated	*37*	*37*
Set	*0*	*0*
Classification	*0*	*0*
Datetype	*0*	*0*
Timetype	*0*	*0*
String	*224*	*1964*
Open	*0*	*-*
Total in data model	*4788*	*18 439*

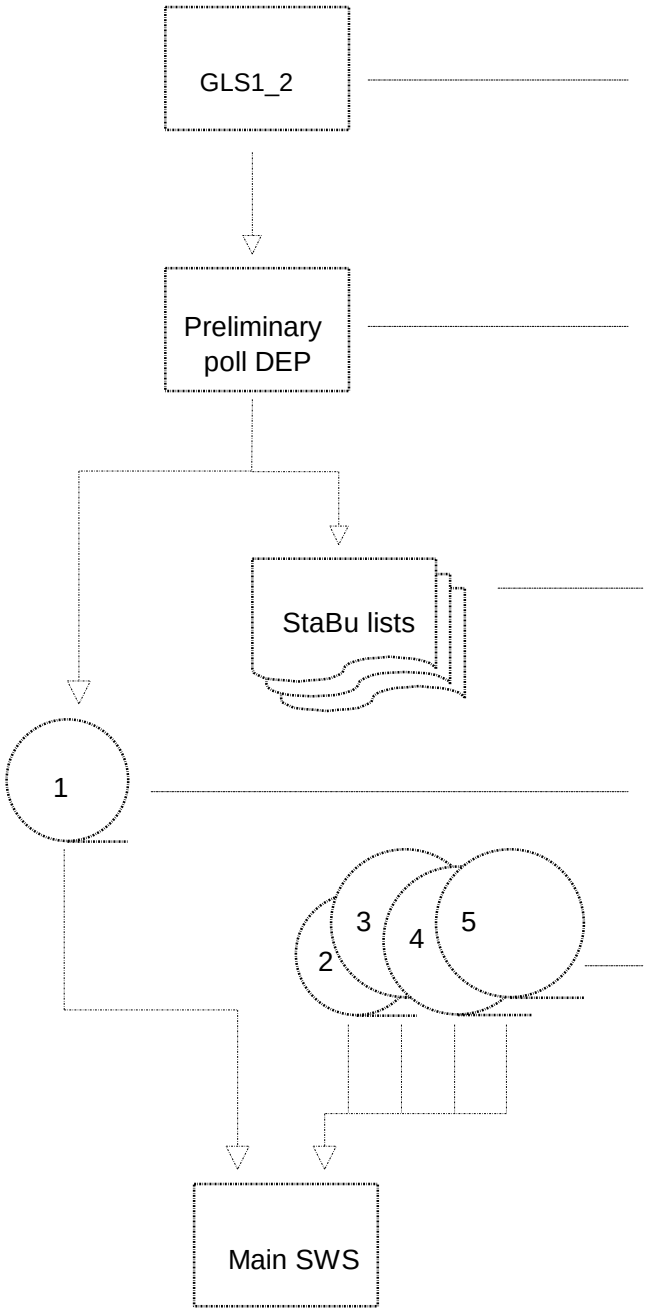
Fig. 2 : Survey of salary and wage structure, data model (schematic)



Legend

	Coding list	Externals	Function
	Collective bargaining agreement text	VORLEIT	Checks
❷	Trade/industry (NACE)	WZLEIT	Coding
❸	Collective bargaining key file	TVOB93	Checks
❹	Deductions group (wage earners)	SCHL_ARB	Checks
❺	Deduction group (salaried employees)	SCHL_ANG	Checks

Fig. 3 : Survey of salary an wage structure, data flow chart



BLAISE data acquisition program for the SWS preliminary poll

The SWS preliminary poll sheet are registered on three PCs

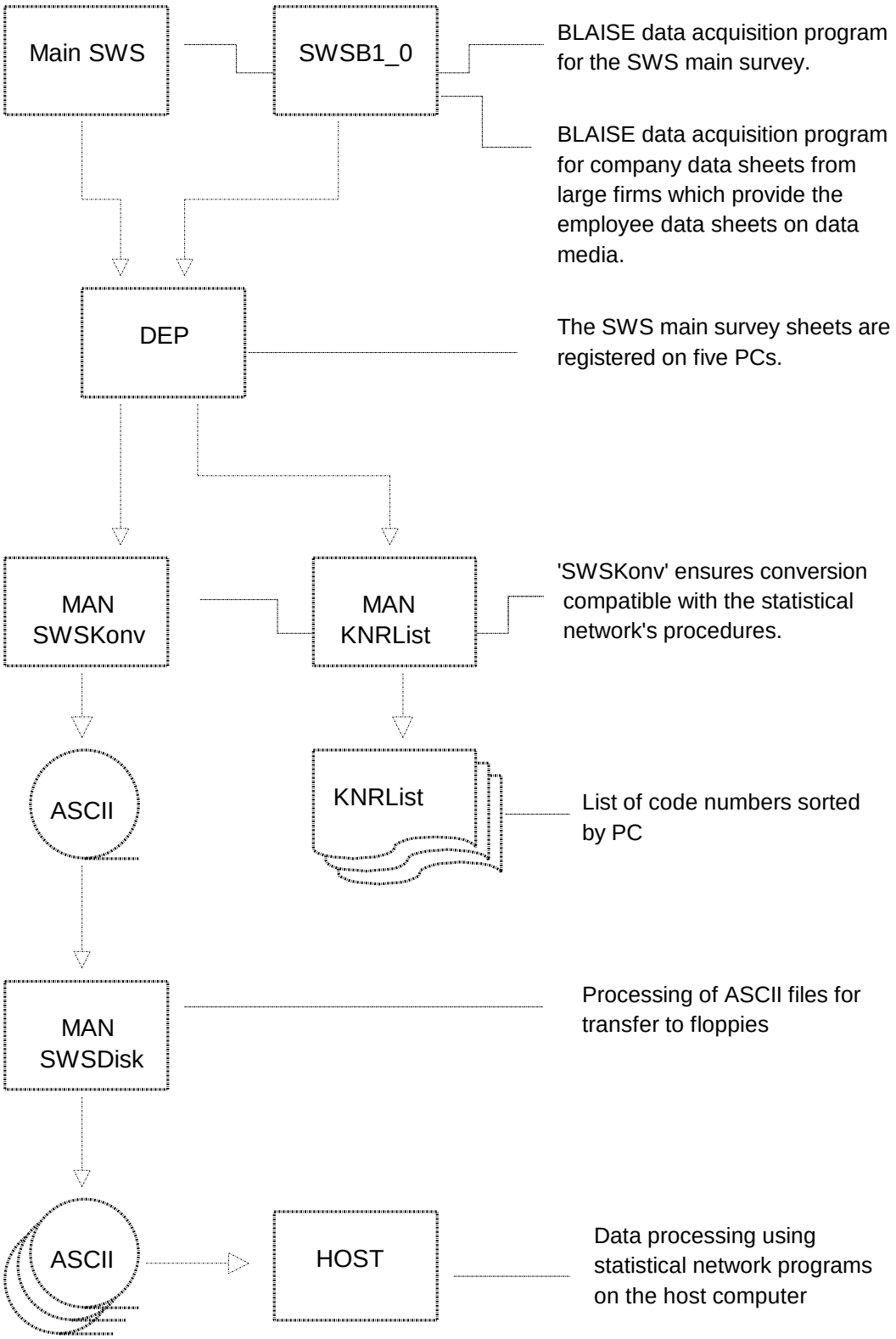
Selected data from the preliminary poll is generated in PRINT format using MANIPULA and sent to the federal statistical office (StaBu).

1) 'VORLEIT': the fields collective bargaining agreement text 1, 2, 3, and 4 from the preliminary poll are integrated as an external file in the main SWS.

In addition to this four files are provided by the federal statistical office which are inserted in the main SWS as external files:

2) 'WZLEIT': file relating to trade/industry system (NACE);
3) 'TVOB93': file relating to collective bargaining agreements (old version); replaced by 'TVOB95' in the course of the survey;
4) 'SCHL_ARB': job coding list for wage earners;
5) 'SCHL_ANG': job coding list for salaried employees

... Fig. 3


Main SWS
SWSB1_0
BLAISE data acquisition program for the SWS main survey.
BLAISE data acquisition program for company data sheets from large firms which provide the employee data sheets on data media.
DEP
The SWS main survey sheets are registered on five PCs.
MAN SWSKonv
MAN KNRList
'SWSKonv' ensures conversion compatible with the statistical network's procedures.
ASCII
KNRList
List of code numbers sorted by PC
MAN SWSDisk
Processing of ASCII files for transfer to floppies
ASCII
HOST
Data processing using statistical network programs on the host computer

The preliminary poll

According to statistical network planning, data acquisition in the states was not to commence until the main survey. The data acquired in the preliminary poll was to be sent in to the federal statistical office in the form of lists for central compilation of a collective bargaining agreement key table. Already at this stage, LDS NRW used a BLAISE data registration program (GLS1_2). This facilitated the achievement of two objectives: firstly, to print out the data acquired and send it to the federal statistical office using a standard paper size suitable for clear presentation of information; secondly, to compile a key file (VORLEIT) specific to North Rhine Westphalia which, taken together with the 1993 version of the collective bargaining agreement key tape, would assure an early start for processing of the main survey data.

The main survey

The data entry program was configured in such a way that the necessary support for validation and coding was provided by a number of externals. Unfortunately the latest collective bargaining agreement key file, which was to supersede the older version, was not provided by the statistical network until data registration was already at an advanced stage. The effects of this delay on error frequency will be referred to later.

The enormous quantity of data from each firm quickly proved to be a problem. At the beginning of data registration work was mostly carried out using PCs with DX386 processors. When data records reached maximum size (e.g. for a firm with one company data sheet, approx. 50 wage earner sheets and approx. 50 salaried employee sheets) the PCs frequently crashed. Subsequent reorganisation runs (MONITOR) occasionally revealed that data had been lost. The systems were not stabilised until the switch was made to PCs with DX486 processors and 8 MB RAM and the number of employee sheets was restricted to 35 per firm.

The changeover from conventional to interactive processing with BLAISE necessitated the creation of an interface to transfer data in a compatible format to the host computer. The federal statistical office’s data record descriptions for SWS served as a basis for the MANIPULA program SWSKonv and ensured a smooth link with the statistical network’s standard program. The conversion procedures required considerably more time than originally estimated. For various reasons it was not possible to network the PCs and the data had to be transferred using floppies. Although the MANIPULA Tools (SWSDisk) proved very useful in splitting the data into manageable chunks, the length of time required was nonetheless unsatisfactory.

5. Results

Having considered the organizational processes and program structures, the question of the benefits of integration and the conclusions that can be drawn is naturally of great interest. This can be assessed by comparing the latest survey of salary and wage structure with that of 1990.

The present article was written approx. 5 months before the planned date for completion of processing and can accordingly provide no final assessment of all the work procedures and effects involved in the use of BLAISE. However, a few initial comments can be made on the basis of findings so far.

Any meaningful interpretation of a comparison of the 1990 and 1995 surveys should, however, be prefaced by one or two remarks on method.

- *The catalogue of statistical characteristics was extended by a further three characteristics.*
- *The work groups were reorganized.*
- *The preliminary poll for the survey of salary and wage structure was also carried out using BLAISE.*
- *Incoming checks and response were effected with computer support.*
- *In 1990 there was a delay of several months before the batch validation programs could be made available by the statistical network and manual validation was accordingly much more extensive than planned.*

The following tables provide an initial overview of the scope, expenditure of time and effort and error frequency of the two surveys.[6]

Tab. 2 : Scope of the survey in terms of the number of firms polled and number of employees

Year	*Firms*	*Employee records*	*Information received on data media in %*	
			Firms	*Records*
1990	*3,000*	*135,000*	*14.3*	*27.4*
1995	*4,000*	*155,000*	*13.9*	*27.7*

[6] *Our thanks to Rolf Schmidt for his contributions and comments.*

Tab. 3 : Expenditure of time and effort and savings

Procedures	*Man months required*		*Savings in man months*
	1990	*1995*	
Polling, processing, evaluation	*290*	*240*[7]	*50*
Registration of data	*10*	*-*	*10*
Support of statistical network programs	*2*	*2*	*-*
BLAISE programming	*-*	*10*	*-10*
PC care and maintenance	*-*	*1*	*-1*
Total	*302*	*253*	*49*

Tab. 4 : Error frequency in first validation batch run (statistical network program)

Year	*Firms checked*	*Errors*	
		Total	*Per firm*
1990	*2,634*	*481,000*	*170*
1995	*3,141*	*162,000*	*52*

The scope of the 1995 survey is just about 15% greater than in 1990. As the proportion of information received on data media remained constant, the number of cases that had to be processed manually was correspondingly larger. This must be taken into consideration when looking at Table 3. Despite the fact that there was more data to process in 1995 the work required for polling and processing was cut by 50 man months. The savings in the central data input were, however, made up for by the work required for BLAISE program development. If the salary and wage structure survey were to be carried out on a continuous basis the programming required for 1996 would not need to be repeated in the subsequent years. Only one or two man months would have to be allowed

[7] *Estimated on the basis of present state of present level of working process until now.*

for maintenance and any necessary adaptation of the BLAISE programs and savings in data acquisition would become really appreciable.

Taking the 1990 figures as a basis, the man months required for the more extensive survey carried out in 1995 would amount to around 330, ie. 27% more than was actually needed. Even when the above-mentioned methodological aspects are taken into account and the effective savings are reduced by a few percent, the benefits of BLAISE are obvious. The cost of buying the necessary PC hardware is hardly significant when compared to the savings in manpower.

This also becomes clear when one looks at the quality of the data undergoing validation. The criterion applied is the number of errors detected in the first run of the statistical network program on the host computer. Any number of errors can be registered per data record. The absolute number of errors was cut to a third and the number of errors per firm to 30%. That a lower error quota results in faster and less costly correction must be clear to anyone.

It is nonetheless surprising that the error frequency was so high in spite of BLAISE. This is partly to do with the organization of data in the process as a whole, because certain things cannot be checked until all the data and preliminary information has been brought together. For example, initial studies show that around 30,000 errors are attributable to the fact that an important coding list[8] was provided by the statistical network so late that a large part of the material could not be compared with it during input. A not inconsiderable number of the errors are soft errors that are recorded for reasons of quality assurance but need not result in corrections in all cases. Account must also be taken of the fact that not all the recognized and conceivable validation checks were implemented in BLAISE so as not to delay sending out the programs to the statistical departments and possibly jeopardize the schedule. For the same reason things concerning validation that were learned in the course of processing the data could be included in the batch programs but not in BLAISE. It is only logical that this resulted in an increased number of error messages.

The bottom line is that PC/host computer integration based on in-house LDS BLAISE programs and statistical network programs has proven its merit. The goals that had been set were achieved in full. The survey was carried out at low cost and a high standard of quality while preserving the methodological uniformity of practice of the statistical network. These impressive results mean that the procedure will be applied to other surveys. Programming with BLAISE can make an efficient contribution to promoting a smooth changeover from conventional batch oriented processing to modern, metadata driven interactive processing of statistics.

Summary

[8] *ie. the so-called "collective bargaining agreement key file".*

For many years official statistics in Germany have been compiled by a network consisting of the federal and state statistical offices. The necessary data processing software is developed jointly and used by all. Today's demands on the statistics production process have given rise to projects aimed at the modernization of data processing as used in the field of statistics. The present article describes how conventional and modern procedures have been combined in the compilation of official statistics in North Rhine Westphalia so as to cut costs while preserving standardized methods within the statistical network. The technical basis for this is the use of PCs and BLAISE III.