LWR: An integrated software system for the household expenditure survey (Continuous economic computations [German abbreviation: LWR])[1]

Thomas Pricking, Landesamt für Datenverarbeitung und Statistik (Federal State Bureau of Data Processing and Statistics) North Rhine-Westphalia (LDS NRW), Germany

1. Preamble

Perusal of the documentation created following the Blaise Users Conferences quickly discloses three extremely comprehensive articles on the subject of *Household Expenditure Surveys*. The central emphasis of this observation varies in this context from the use of encoding methods (ROESSINGH et al., 1993), via comparative assessment of the PAPI and CATI surveying procedures (MANNERS, 1995), up to and including the use of Blaise III as software for the integration of a range of different surveys (MANNERS et al., 1997). SCHWAMB and WEIN (1997) also examined this subject. The question arises in view of these publications of whether there is anything new to report at all, and if so, what can be said on the subject of *Household Expenditure Surveys* which is new, from the point of view of Blaise programmers and users.

The term "new" can be viewed in this context from a number of different angles: On the one hand, there is the technical viewpoint, as to whether new potentials of the Blaise development tools are exploited, for example. I intend to take only the cursory look at this subject. On the other hand, the taking of particular account of users with their special needs and tasks in the development of a software package can result in a new mode of observation. This is precisely the focus of this essay: Is it possible using Blaise to provide specialist statistical departments with support in such a way that account can be taken of complex conditions and extremely heterogeneous requirements? The basic statistical fundamentals are firstly described, in order to render planning activities and their implementation comprehensible in the form of a **experience report**; the basic outlines of the statisticians‘ list of requirements is then examined, and the solution finally presented.

2. Statistical fundamentals

The statistics of ongoing household computations have been performed at regular intervals in Germany since 1950. Throughout Germany, 1000 households initially took part; these were joined after re-unification as from 1993 by a further 1000 households on a voluntary basis, categorized into three household types: Employees' households with average income levels and government officers' and salaried staff households with higher income levels, each consisting of four persons, and two-person households in the form of pensioners and recipients of social assistance. These households kept a monthly household ledger in which all incomings and expenditure were recorded in minute detail. The households generally remained in the random sample for a number of years, in order to ensure the most continuous flow of data possible. Data preparation in the statistical bureaus was accomplished using traditional means: The dossiers received were checked manually, and allocation of the products acquired by the household into the books was performed by hand. Once the central data acquisition elements had entered the comprehensive survey data into a mainframe computer, the data was ordered in batch runs and manual intermediate steps. The survey was administered by means of fling cards.

[1] This article and the corresponding presentation are dedicated to Bernward Hausmann, who continuously promoted the use of Blaise throughout his period of service with LDS NRW.

This procedure, which was an accepted practice in the 1960s and 1970s, had fallen, by the 1990s, at the latest, into technical and organizational disuse: The not inconsiderable burden placed on the households led to a drop in their willingness to participate, which for its part produced a decline in representative quality. The period between the performance of the survey and publication was considered excessively long, and the processing phase to be of poor effectiveness and efficiency.

For this reason, the German Office of Statistics decided in 1994 to overcome these difficulties by means of a new methodological concept. The following restructuring measures were to be implemented:

- New survey ledgers were intended to make it easier for the households to provide answers to the comprehensive list of questions.
- A rolling quarterly panel was created, in order to reduce the burden on the participating households. The random sample was now to consist of 6000 households, each household needed to respond only once each quarter, however. This provision was intended to retain the legally specified scope of random sampling (2000 households) and, at the same time, achieve a broader scatter of households.
- The constitution of the random sample was to be placed on a new basis, and, therefore, to achieve greater representative quality, via orientation around other socio-demographic facts derived from the micro-census.
- Technical support was to be placed the completely new foundation. The batch-based procedures were to be superseded by dialog-based solutions, with comprehensive support using computer-based means targeted for all working processes.

The feasibility of the new conception was examined in a pilot study in 1996 and 1997 (see CHLUMSKY et al., 1997 and GERTKEMPER, 1997). In view of the positive results obtained, LDS NRW was awarded in 1998 a commission by the Statistische Verbund for the drafting of the programs necessary for the new procedures.

3. Requirements on the computer system

A data-processing project is generally preceded by an assignment analysis, in order to determine the subsequent users‘ technical and organizational needs. Two specialists, from our own statistical bureau, on the one hand, and from the Federal Bureau of Statistics, on the other hand, were available as contacts for the four-person development team. Personal know-how acquired in the context of test surveying, was not available for use, since both bureaus had implemented a change of personnel as a result of internal restructuring provisions in the field of economic computation. In addition, it quickly became apparent that the programs produced for the test survey had largely not subsequently been used, for a range of different reasons. In short: It was necessary to develop and program a completely new data-processing system. Luckily, the concepts of the technical representatives complemented each other ideally: The Federal Bureau of Statistics placed particular emphasis on the taking of random samples, plausibility checking and drafting of results, whereas the manner of observation of the LDS representative concentrated on survey organization and on the practical usability of the programs. Nine modules which would need to be created became apparent relatively quickly:

a) Address management for interested and participating households;
b) Random-sampling takings;
c) Survey management;
d) Acquisition and plausibility checking of the statistical data;
e) Setting-up and upkeep of a classification system (COICOP);
f) Budgeting of the households‘ income and expenditure;
g) Classification of the households;
h) Exportation of the ordered data for determination of weighting factors;
i) Re-importation of the weighting factors and exportation for tabularizarion.

The LDS NRW had defined for the test survey a base module **address management**. It was, it is true, possible to integrate this module into the basic structure, but it needed to be augmented with significant elements. It was necessary to integrate both an importation program and a checking for duplication, since use was to be made for the survey of existing addresses of the households which had already made up the reporting group or had participated in random sampling of income and consumption and, in addition, further households were to be recruited by means of advertising provisions. In order to simplify searching in the address basis, further, secondary keys were generated in addition to the household number constituting the primary key. The post-code directory and bank sorting code directory were, of course, integrated as coding files, in addition to a range of export formats, in order to facilitate further utilization of the address data in the form of multiple letters, label stickers, etc..

Although it was possible to implement these provisions without further difficulty using the standard Blaise solutions, it became apparent during programming of the **random sampling procedure** that the principle of data storage using Blaise necessitated redesign of the technical concept. The random sample forming the basis for continuous economic computations is generated in the form of a quota random sample. The quotas are based on combinations of the following features:

Household type	Net household income	Earning
Single-person household	less than 1,000 DM	Yes
Marriage/common-law marriages, no children	1,000 – 2,500 DM	
Single-parent family	2,500 – 5,000 DM	No
Marriage/common-law marriage, with children	5,000 – 7,000 DM	
Other household type	7,000 or more DM	

Combination of the feature characteristics produces a maximum of fifty layers, of which 15 are combined with other layers, however, with the result that there are 35 layers to be assigned. The quota plan defines the number of households per layer as a function of population number and structure of the individual states, there being a significant breadth of span: Of the total of 6000 households, for example, 1050 households are located in NRW; the largest layer contains 90 households, whereas there are others with less than five households. The households in each layer must be uniformly distributed within the quarter, in order to avoid seasonal "clumping" effects. Households with similar classification features may be allocated by way of substitution in accordance with a detailed depiction specification if the quota plan cannot be completely fulfilled.

It appears an obvious step to create a number of Blaise files for this purpose, in order to be able to control the procedure with a simple algorithm: As soon as a household is allocated to a layer as a result of its structural features, and as soon as this has been assured, a corresponding change is implemented in the other files and must also be stored, for reasons of consistency. At precisely this point, the Blaise file system is unable to provide any support: It is not permissible to have a number of files open simultaneously or reading **and** writing, as is possible without any difficulty in all PC data-bases. Since API interfaces were still to come at the time of Blaise III, but work was, on the other hand, to be performed using the Blaise system, it was necessary to find alternative solutions. Ultimately, it proved possible to convert the complex random-sampling procedure into a sequence of steps which, from the basic "mass", the households, the features of which accorded precisely with the sorting structure, were allocated for the random sample. All the households not selected were classified as so-called "standby" households. The random sample is filled out from this stock by way of substitution in a second step. Following this allocation the households are subdivided to the so-called "monthly waves" on the basis of the quota plan. Records document these procedures in great detail, with the result that the specialist statisticians can obtain a precise overview of fulfillment of the quota specifications at any time. Administration of the participating households is an essential component of **survey management**. The households cooperate voluntarily in this comprehensive statistical exercise. Careful upkeep of the data

stocks thus obtained is equally significant for the continuity of the reporting group and for rapid and flexible reaction in case of queries, problems and proposals for changes. A number of tasks therefore arise for survey management:

- The addresses must be up-to-date;
- The documentation for the participating households must be prepared and posted;
- It must be possible to directly register any households which wish to take a break for a quote, keep the household ledger in another month, or leave the survey, and to draw corresponding consequences for posting of documents, etc.;
- Where a household has submitted its documentation, a corresponding note of receipt must be implemented, in order that unnecessary inquiries are avoided;
- It should, on the other hand, be possible to send reminders for documents not yet received.
- The participating households receive a modest payment. It must be possible to pay out these amounts immediately and by means of non-receipted transactions;
- It must be possible to draft an up-to-date status report on the status of preparation at any time. A range of business statistics need to be implemented for this purpose.

The backbone for all these activities is the so-called "Household file" (Figure 1: LWRHHD), which constitutes the central control element. In addition, it is also automatically used in the background for other important working operations within continuous economic computations. The data-model fields necessary for this purpose are kept hidden from the user in the dialogue.

The great extent to which flexibility is required of survey management can be illustrated using the example of payment of the premium. It was necessary, since each bureau of statistics wishes to set different background conditions for the payment, i.e., make payment of the payment dependant on receipt of data and processing status, to program a dynamically configurable check condition:

Receipt of documentation	Logical link	Processing status
Yes	and	Undergoing processing
		Enquiry
		Editing completed
No	or	Budgeting completed
		Annual accounting completed
		Forwarded to Federal Bureau of Statistics

The corresponding parameters are set in a Blaise file (INIT), read in in the context of a manipulation operation, and control payment of the premiums.

Survey management is concerned not only with cooperation with the households. Working procedures within the bureaus of statistics must also be controlled. This is reflected in the form of staff administration governing the various editing authorizations of the various users: The Supervisor has access to all functions and is therefore the basic "boss" of the operation, whereas the "standard member of staff" can perform only specific operations, such as the input of questionnaires, and elimination of errors, for example. The Bureaus of Statistics also employ so-called "home workers", who are furnished at home with technical bureau equipment and register and edit the household ledgers at home. Work "portions" are generated by means of survey management for this purpose; this operation also includes the effective anonymization of address information necessary for data security reasons.

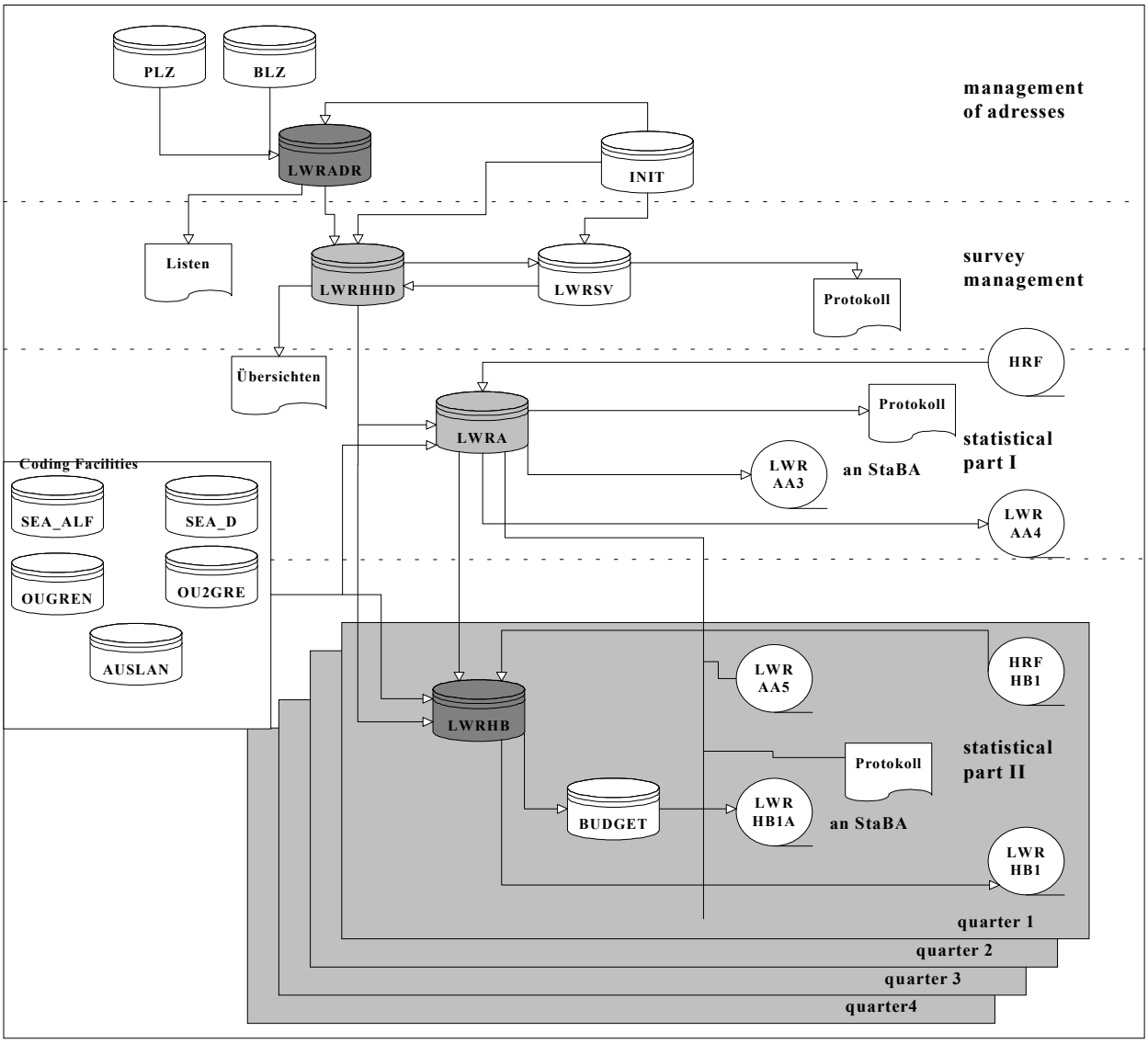

Figure 1: Schematic overview of datamodels in the integrated survey systems.

The central technical element of this application is made up, of course, by the modules for **acquisition and plausibility checking** of the statistical information. Continuous Economic Computations consist of two parts; in the first part, the so-called ***General Information***, information on the social status of the members of the household, their employment, information on the size of the household and on net income, domestic situation and availability of consumer goods is obtained. The twelve-page booklet is sent to the households at the beginning of the year and is required back by the end of January.

This questionnaire does not make any special demands in technical terms. Since the list of consumer goods available is modified each year, a flexible access logic which selects and displays the goods concerned in the survey year from the totality of possible goods has been created for this purpose using an *external file*.

The second part of the survey is made up of the so-called ***household ledgers***, in which the households record in detail all incomings and expenditure, as well as all changes which have occurred in the household since the last reporting period. Only a few figures are needed to illustrate the statistical and

technical complexity of this section of the survey: The technical department has drafted 200 error specifications in which all fields occurring in practice are involved. Since it is possible to register up to ten persons in each household, and dependency relationships exist between the individual members of the household (parents – children, earners – dependent persons, etc.), the logical necessities and the associated program input can easy be imagined.

The sequencing control of the surveying process itself constitutes an initial challenge:

- Basic socio-demographic features must be extrapolated from the *General Information* element of the survey in such a way that they can be used for plausibility checking of the data in the household ledgers. This operation must be controlled as a function of the data quality achieved and, in addition, it must be possible to repeat it as often as needed and take all quarters into account without endangering data integrity.
- Extrapolation factors applicable for the whole of Germany need to be determined in order to permit evaluation of the data. For this purpose, the technical concept includes central provision by all state bureaus of the plausible data stock for each quarter. The Federal Bureau of Statistics determines the extrapolation factors for the entire random sample and sends it back to the individual states on a case-linked basis. The states, for their part, enter the factors into the data record and are then able to perform evaluations. It must be ensured in the case of this procedure that further modification of the data after export is no longer possible; each and every modification in, for instance, social status would have implications for the extrapolation. For this reason, a concept which "seals" the data, so to speak, at export has been developed and implemented. Such "marked" data records can be viewed only, and not edited in any way. Whereas no system-side solutions for these needs were available in Blaise III, and separate views and browsers needed to be programmed with no small degree of difficulty, it finally became possible to implement an up-to-date access logic system by means of the READ ONLY parameter as from the Windows version (4.3).
- The "sealing" procedure, which provides the greatest possible protection for the data against erroneous editing, has now proven in individual cases to be a hindrance in error elimination. Errors resulting from inadequate specifications were discovered only at the evaluation stage. Thanks to the "sealing" process, however, access to the data had been disabled. It was necessary to introduce a special procedure for cancellation of "sealing" in order to provide a remedy in such problem cases, however. For security reasons, this is available only to the system manager.

The programming of the household ledgers opened up a new dimension to the team of developers: Each data record is made up of more than 7200 fields and takes 35 KB of storage capacity. This volume is the result not only of the extensive person-relevant data on incomings but also, and in particular, of the portion of the household ledgers in which daily expenditure is recorded. Nineteen pages with 600 lines are available to the households for recording in this case, whereby not only the precise designation the article and the amount are recorded, but also whether the expenditure was in cash or not, in DM or €, in Germany or abroad. An obvious step would be to convert these questions to a table, into which the highly efficient and powerful Blaise coding tools should be integrated. Annoying problems occurred during implementation and test operation, however:

- Since special checks were to be made within the various expenditures – if, for example, a tenant has stated expenditure for rent, this was to be accomplished in a single table spread over 600 lines. The technical people also wanted an error checking system for adherence to upper and lower limits for each expenditure, in addition. Building of the table on the monitor screen was significantly too slow, despite the fact that powerful PCs were available in 1999. Waiting times of up to one minute were not unusual. In order to accelerate the procedure, the range-limit check was firstly redesigned in such a way that it was only applied upon express request. It has now been removed completely. Survey practice also indicates that scarcely more than 300 to 400 entries are normally necessary. The user can now therefore define the number of lines which he intends to enter in a variable before loading the

table. In addition, he can also close the table prematurely using an abort condition (End or Code 9999999, see Figure 2).

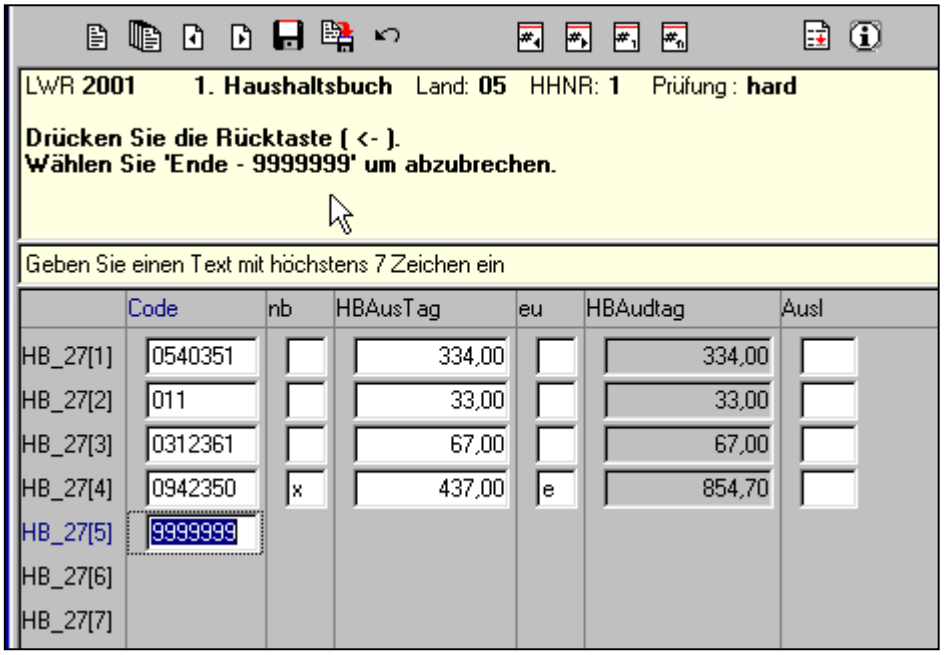

Figure 2: Enter code 9999999 to leave table.

- Allocation of daily expenditure is based on SEA (Systematic Index of Income and Expenditure by Private Households) classification, which is derived from COICOP/HBS systems (Classification of Individual Consumption by Purpose/Household Budget Surveys). It was intended to provide this classification to the user both in the form of a hierarchical "search tree" and in the form of an alphabetically sorted list complete with trigram coding. During conversion of the SEA classification to a Blaise coding file, however, it became apparent that the hierarchical search mechanism does not function correctly under Blaise III and cannot be used. The SEA code contains a two-digit code (01, 02, ... 11, 12) as its first hierarchical level. The order of the codes is reversed (e.g. 02 after code 12) as a result of the coding tool's sorting method, since the leading zero is obviously not taken into account. The SEA classification texts were analyzed, dismantled and re-processed in order to ensure the greatest possible hit rate using alphabetical searching, with the result that they are now available with a short text and an explanatory long text for searching (Figure 3). This procedure has proven extremely worthwhile; approx. 90 % of the codes were found immediately when the goods designation was entered. The hit rate increases by five percent at changeover to trigram mode.[2] Originally, it was to have been possible to augment the coding files at any time, in order to permit inclusion of new consumer goods and also of synonyms, for example. Sine extremely significant performance problems occurred with multiple users in network operation in the necessary read/write mode, a successful change was made to read-only mode. In addition, the users can transfer the coding files to the local machines and access them there directly if the LAN should prove too slow. All the innovations and augmentations for classification have now been centrally compiled and technically revised and sent to the Statistische Verbund in the form of directly usable coding files. LDS NRW has transferred its own application for upkeep of the SEA register to the Statistisches Bundesamt for this purpose.

[2] It would be extremely practical if the search strings were retained in the entry box at a change of search mode under B4W, as was the case under Blaise III!

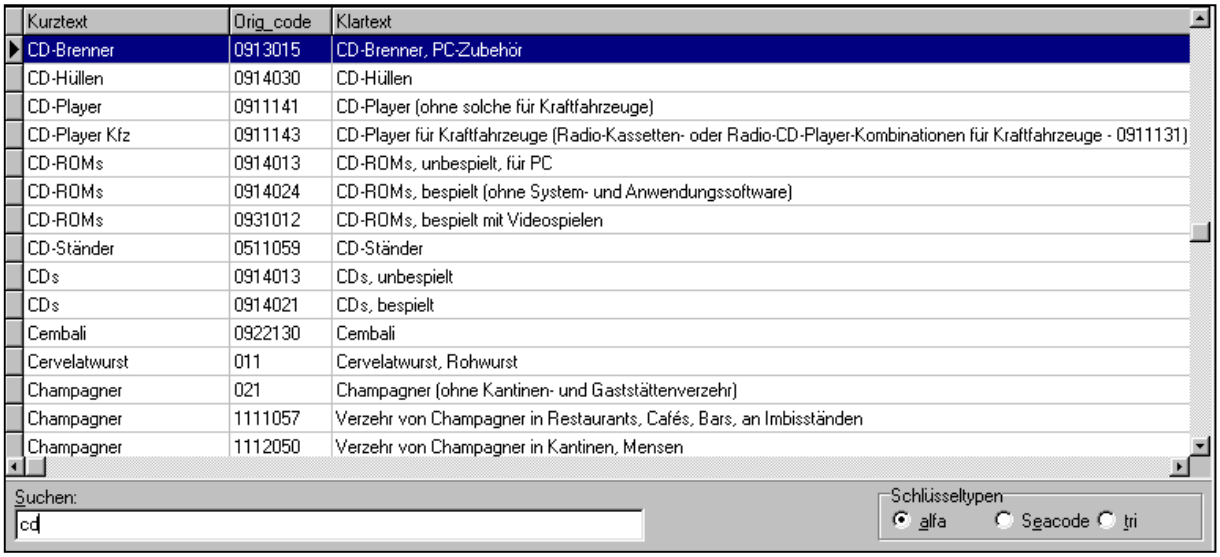

Figure. 3: Alphabetical coding in LWR2001 (Table-heading, short text, SEA code, long text)

The methods for **budgeting, classification** of households and **exportation** of ordered data and for **re-importation** of the weighting factors are not examined in more detail here on technical criteria. The technical support necessary for these processing steps has been implemented in comprehensive Manipula programs. Despite the fact that only 1050 households need to be processed in each quarter in NRW, these programs in some cases require extensive running times, due to their technical complexity and the voluminous data models involved. Luckily, the changeover to B4W produced unexpected performance enhancements for data export: These programs are now completed in a few minutes, whereas they needed to run for a number of hours under Blaise III. This is, unfortunately, not also the case for the subsequent plausibility runs necessary (C/S, see below), with the result that dialogue-mode operation is not possible during this period.

4. General aspects

The complexity of the application as a whole can with certainty be illustrated by the fact that more than twenty in some cases extremely voluminous data-models needed to be defined, and just on sixty manipula set-ups generated. The structure chart in Figure 4 also illustrates this. The team of developers managed to model and implement all working procedures in technical programming terms in such a way that they can be started within the application. In addition, the changeover from Blaise III to B4W permitted the integration of the MS Office package. Evaluations and graphic presentations of the business statistics can be called up by the users independently without difficulty. The work input for development of the concept, implementation, tests and introduction was high, however. A total of more than six developer years was invested in all[3]. Acceptance by users at all the statistical bureaus is good; it has been determined in the course of a long-term observation that all requirements have been implemented in a manner appropriate to everyday practice. Despite this note of acknowledgement, a number of criticisms are nonetheless appropriate on a technical and organizational basis.

[3] Particular thanks to B. Wlodkowski, F. Merks, A. Hansen, M. Broose and the Blaise support team from CBS NL.

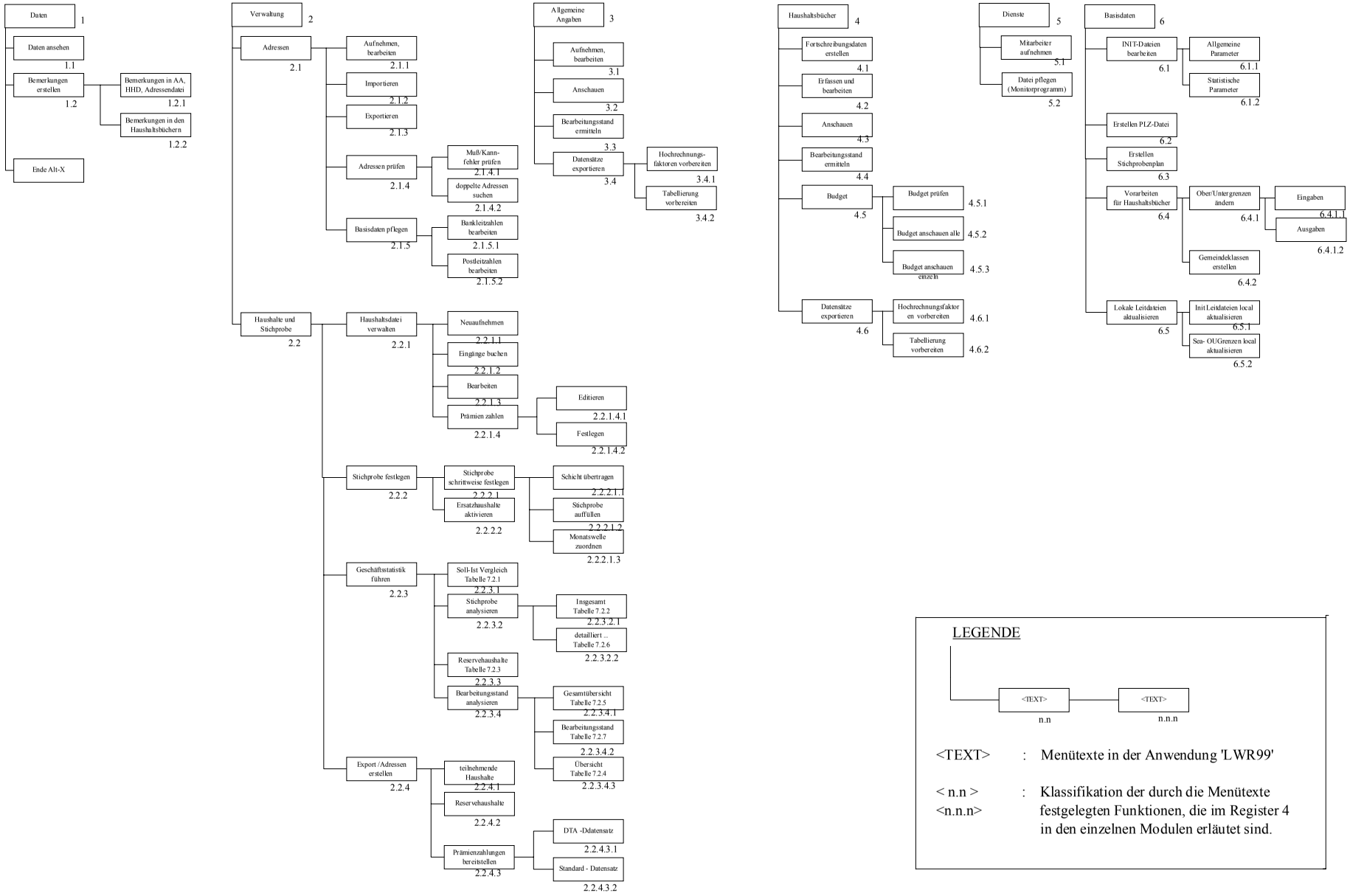

Fig. 4: Functional structure of LWR2001

Blaise's DEP module in principle makes it possible to select various modes for data input and for the corresponding monitor-screen presentation. Only the extremely inconspicuous error counters could be used for error identification in edit mode, however, since development started at the relevant time using Blaise III. Mandatory errors cannot be suppressed in purely interview mode to permit their subsequent correction after enquiry, however. It was necessary to search for other methods in order to permit rapid input of data in interview mode with complete error signalization and the option of suppressing it. A variable with the *check/signal* (C/S) check specification is pre-assigned in the INIT file mentioned above. The overall routing of the data model is structured in such a way that all checks are integrated into a program sequence which is tripped by the check specification (either "checks" or "signals"); see Figure 5.

```
{Check or signal is active until new command}

IF exini.hb_prufe[q] = muss THEN              {muss: var. in INIT.DBD; muss = hard error}
      CHECK
ELSE
      SIGNAL
ENDIF

IF (nb = EMPTY) AND (eu <> EMPTY) THEN
                  exini.jahr >= 2002
 INVOLVING(eu, nb)
                  "No EURO-cash before 2002!"
ENDIF

IF ((Code ='2099201') OR (Code = '2099301')) AND (nb = '5') THEN
      eu = EMPTY
            "[S19B*] The "Euro" must not be assigned with "x" or "e" for entry of the number
            of garages and parking spaces rented"
ENDIF
```

Figure. 5: Example of external dynamic checking

In order to be able to work quickly, users generally suppress all errors. Since it is necessary to ensure that only error-free material is used for further purposes, however, a concluding check run, in which all mandatory errors are determined, must be performed. This procedure also makes use of the solution described: "Check" is automatically selected in the INIT file and all records are checked in the batch. Simple error statistics are drafted simultaneously. Unfortunately, these runs block the system for a long time – a period of three to four hours is not uncommon, with the result that they are usually performed overnight.

These performance problems in a number of processes in this application have been criticized by users. Around twelve members of staff normally use this application in NRW. The team is doubled at peak periods, however, with the result that all disruptions have a long-term effect on the data preparation process. Improvements in the running of plausibility checks in large tables and in the handling of transactions in the Blaise running-time system would be extremely welcome here.

A problem unsolved up to now is constituted by occasional losses of data: The application crashes on the local workstation during storage of a data record, and the data input for a household, which may in some cases have taken up to two or three hours of work, is partially and in some cases completely lost. Regrettably, no regularities are discernible, signifying that defect analysis and trouble-shooting are difficult. Since the programs used are obviously syntactically correct and also function in principle

correctly, the Blaise system can apparently be eliminated as the cause of the problem. The LDS has made intensive efforts to determine other possible causes of these crashes. The physical structure of the network was first examined. The problems continued to occur even after relocation to a new, and technically optimally equipped, office building. The suggestions provided by CBS of modifying network parameters also failed to produce any radical improvement. At present, attempts are being made by means of a precise addressing system for each PC used to identify possible local features of specific machines as the causal element.

A concluding remark relates to the change of version from Blaise III to B4W. The announcement that the DOS programs would be converted to Windows at the push of a button produced skepticism in advance on the part of the developers. Even the Blaise team was itself a priori convinced that the changeover could be accomplished without difficulty. This skepticism was, it is true, dispelled after the first few successful attempts, but it remains a fact that it is necessary to invest more work than had initially been assumed. This was the result, on the one hand, of improvements which derived from the new potentials (linking to MS Office) and, on the other hand, all the modules had to be re-converted, checked and integrated into the application. During this donkey work, difficulties with the standard conversion routine (B4W, Version 4.2/4.3) for conversion from ASCII to ANSI were just as much of a hindrance as the occasionally incomplete documentation. The support provided by CBS ultimately made it possible to solve the conversion problems well, however.

5. Conclusion

To answer the question initially posed of whether it is possible to draft complex applications for statistical purposes using Blaise, the answer is, of course, "Yes". It has become apparent that it was possible to satisfy the extremely diverse technical requirements practically completely, particularly following conversion to Windows version. The Blaise development environment contains numerous high-power tools for the achievement, on the one hand, of tight intermeshing of computer-based data-processing support and statistical preparation processes while, on the other hand, it is equally possible to implement a survey management system with complete address, household, staff and random-sample administration leaving nothing to be desired on the part of the users. It is therefore indeed worthwhile to further pursue the integrative conception outlined for other statistics, too. It would, however, then be desirable to provide certain modules – such as the address and staff administration modules, for example – i.e., basic "building elements", in a generalized form which can be re-utilized at any time. Blaise's telephone management system is an excellent example of such a procedure. Initial developments in this direction are now becoming apparent in the joint work performed in Germany by the Statistische Verbund; it would be desirable for such modules to be available to all Blaise users – a worthwhile exercise for CBS NL.

References

Chlumsky, J. and Ehling, M. (1997), Grundzüge des künftigen Konzepts der Wirtschaftsrechnungen der privaten Haushalte. In: Wirtschaft und Statistik 7/1997, S. 455-461, Wiesbaden, Statistisches Bundesamt.

Gertkemper, F. (1997), Ergebnisse der Testerhebung zum neuen Konzept der Laufenden Wirtschafts-rechnungen. In: Wirtschaft und Statistik 12/1997, S. 872-878, Wiesbaden, Statistisches Bundesamt.

Manners, T. (1995), The UK Family Expenditure Surrvey: comparison of PAPI and CAPI estimates from a controlled trial using Blaise. In: Essay on Blaise 1993. Proceedings of the Third International Blaise Users Conference, pp. 101-114, Helsinki, Statistics Finland.

Manners, T. and Deacon, K. (1997), An Integrated Household Survey in the UK: The Role of Blaise III in an Experimental Development by the Office For National Statistics. In: Actes de la 4è Conférence Internationale des Utilisateurs de Blaise 1997, pp. 197-211, Paris, Institut National de la Statistique et des Etudes Economiques.

Roessingh, M. and Bethlehem, J. (1993), Trigram coding in the Family Expenditure Survey of th CBS. In: Essay on Blaise 1993. Proceedings of the Second International Blaise Users Conference, pp. 118-132, London, Office of Population Censuses and Surveys.

Schwamb, J. and Wein, E. (1997), Notes on an Integrated Data Management. In: Newsletter 10, pp. 33-35, ed. by International Blaise User Group, London, Office of Population Censuses and Surveys.