COLLECTION OF RAW DATA

TASK FORCE

MEETING Nº 3

4 OCTOBER 2000

Doc. CoRD 026

Developing a system for electronic raw data collection at Statistics Austria

For information
See Work Programme, Item 4 (Doc CoRD015)

Collection of Raw Data Task Force
Meeting No. 3 (4 October 2000)

Developing a system for electronic raw data collection at Statistics Austria

by Wolfgang Koller and Günther Zettl

Remark: This document is an updated version of an *invited paper* that was presented at the UN/ECE "Seminar on integrated statistical information systems and related matters (ISIS 2000)" (Riga, Latvia, 29 – 31 May 2000).

I. History

In 1998 Statistics Austria was experiencing a lot of political pressure (for example, from the Chamber of Commerce and Industry and from the Minister of Economic Affairs) that – in order to lower the burden for the enterprises – software for collecting and transmitting statistical data should be provided to the respondents of business surveys free of charge. Later on, this request became part of the "Federal Statistics Law 2000"[1].

In June 1998 a working group was founded consisting of members of the IT department and of the department for business statistics. Within three weeks we produced the first draft of a requirements analysis for an electronic questionnaire software. In July 1998 Statistics Austria officially agreed to start an electronic raw data collection project which was later called „SDSE – System zur Durchführung statistischer Erhebungen" (system for carrying out statistical surveys). At this time the main focus of our considerations was on the *structural business statistics*, an annual survey with approximately 40 000 respondents.

Statistics Austria did not have the manpower nor the PC programming skills to develop the software in its IT department, therefore we began to prepare a call for tenders. In the course of this work the scope of the project was growing rapidly. In our view, it was not a good idea to develop an electronic questionnaire software specifically for a certain survey, as every change of the survey would require a corresponding adaptation of the program's source code, and it would also be necessary to develop new questionnaire software for every future survey offering the possibility of electronic responses.

So the project goal became a more general one: the core element of the SDSE should be an "electronic questionnaire management system" that could be used for different (economic as well as non-economic) surveys by specifying all survey-specific

[1] § 28 (3): „Auf Wunsch sind den Auskunftspflichtigen die entsprechenden Unterlagen für die Auskunftserteilung auch auf elektronischem Wege kostenlos zur Verfügung zu stellen, soweit dies zweckmäßig und aus fachlichen Gründen vertretbar ist." ("On request, the respective supporting material for electronic responses must be placed at the disposal of the respondents free of charge, as long as this is useful as well as technically justifyable." The comments of the law explain that "respective supporting material" means "mostly software suitable for the preparation, control, and transmission of the necessary information").

information, including questionnaires and validity checks, in XML parameter files. Respondents and intermediaries (third party declarants) such as accountant firms will get this software free of charge to fill in the questionnaires or to import data from their own EDP-systems, to manage the response data and to send the encrypted data in an XML format via e-mail, FTP or mailbox to Statistics Austria. The program is planned to be used by Statistics Austria as well, so that our expert statisticians can view, check and edit incoming data with the same tool that the respondents are using.

In March 1999 we presented our concept to representatives of the Chamber of Commerce and Industry and to several enterprises. Ideas and proposals raised at this meeting were added to the requirements document that we were writing for the call for tenders.

In May 1999 it was decided that instead of *structural business statistics* the monthly *short term statistics* should be the first survey to utilize the new system for electronic raw data collection. Deadlines for a pilot test (October 2000) and full operation (January 2001) were fixed.

At the beginning of July 1999 the international call for tenders (carried out as a two-step negotiation procedure) was published. 15 software development companies answered the call. On August 24 six bidders were selected for the second phase. In addition to the requirements document that was sent to them, we organized an obligatory meeting to supply background information and give them the opportunity to ask further questions.

The second phase of the call for tenders ended on October 18, 1999. Four of the six participating companies submitted concepts for the realization of the SDSE. On November 15 *CSC Servodata* (now called *CSC Austria*) was selected. In close cooperation with Statistics Austria, they immediately started to work on the project. The first major milestone, a detailed requirements analysis, was finished on February 23, 2000.

At the end of May a first prototype of the software, which was already able to dynamically generate a questionnaire based on XML metadata (but of course with reduced functionality), was installed at Statistics Austria so that we could perform first tests. This prototype included an alpha version of the PRODCOM classification component.

On June 27 the current state of the development and some background information were presented to representatives of the Chamber of Commerce and Industry as well as of enterprises which agreed to take part in the pilot test in October.

During the summer months, the IT division of Statistics Austria started to work on the integration of electronic responses for *short term statistics* into existing processing systems. Members of the project team from the business statistics division elaborated the text for the help system of the electronic questionnaire software. We also prepared several marketing activities; for example, at the end of August we submitted a letter with some information about the project to every repondent of the *short term statistics* survey. In addition, we asked the respondents if they intended to use the program (presenting, among others, questions regarding their EDP hardware and software). Within two weeks, 1100 answers were received, and two thirds of these respondents

expressed their interest. We also had a first meeting on the subject of supporting the import interface of our questionnaire software in SAP R/3.

Meanwhile CSC Austria continued the development of the software. A second prototype was finished in the midst of August and the so-called alpha version (with almost all features of the final product implemented) was deployed on September 5.

II. System overview

The SDSE is a software system for electronic raw data collection consisting of three sub-systems (fig. 1).

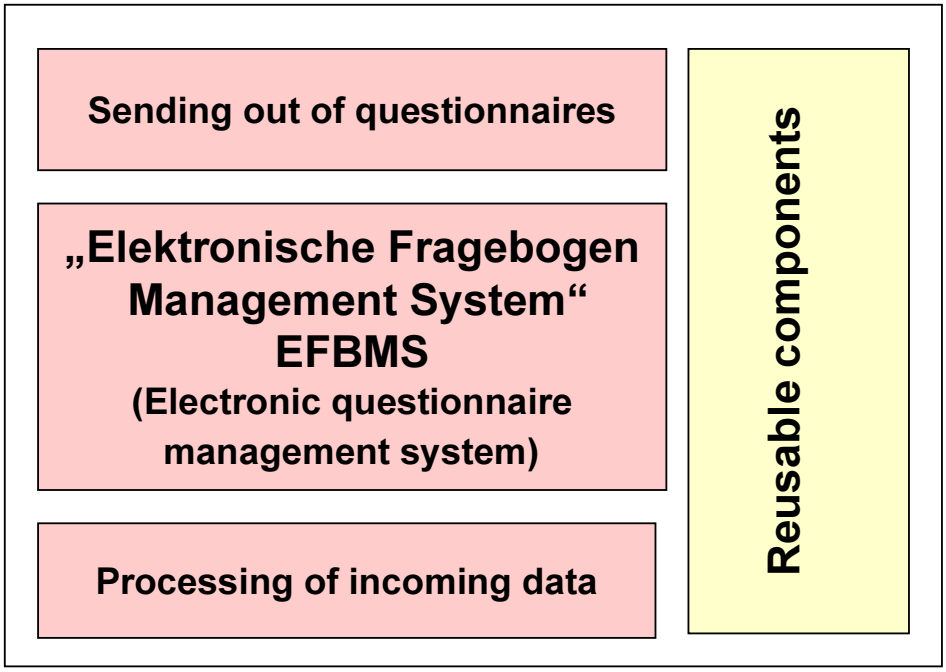

Fig. 1

- The sub-system “Sending out of questionnaires“ consists of software to encrypt and compress all XML parameter files that are necessary for EFBMS (see next item). There will be two types of parameter files:

  - structural data (describing the collector of statistical information, the survey and its versions, the types of observation units and their respective questionnaires (including validity checks), hierarchical relationships between observation unit types, and some more objects)

  - and respondent-specific data (the actual observation units for which the respondent must fill in questionnaires, the actual relationships between them and initialization data that has to be imported into new questionnaires).

  Also part of this sub-system will be a tool for designing questionnaires and for managing structural XML parameters (“EFBMS metadata management system“).

- The EFBMS program ("Elektronisches Fragebogen Management System" – electronic questionnaire management system) is the most important – and most complex – component of the SDSE. On the one hand, it will be put at the disposal of the respondents, so that they can use it for the collection and administration of their statistical declarations as well as for the electronic transmission of the response data to Statistics Austria (and in future, it will possibly be made available also to other institutions using EFBMS for their own surveys), on the other hand, the staff of Statistics Austria should also be able to use it for the viewing and the processing of the transferred data.

- The third sub-system "Processing of incoming data" consists of programs which fetch the statistical declarations from e-mail, FTP and mailbox servers in regular intervals, backup, decode and decompress them and register the arrival of the responses in a database. Then the data are passed on to the responsible organizational unit (fig. 2). The expert statisticians will have an online application to administrate the incoming response data files (tentatively called the "pot application"). For viewing and correcting the contents of a file EFBMS will be used (fig. 3). Finally, the data will be converted and transferred to the mainframe computer where further processing will be the same as for responses originating from paper questionnaires.

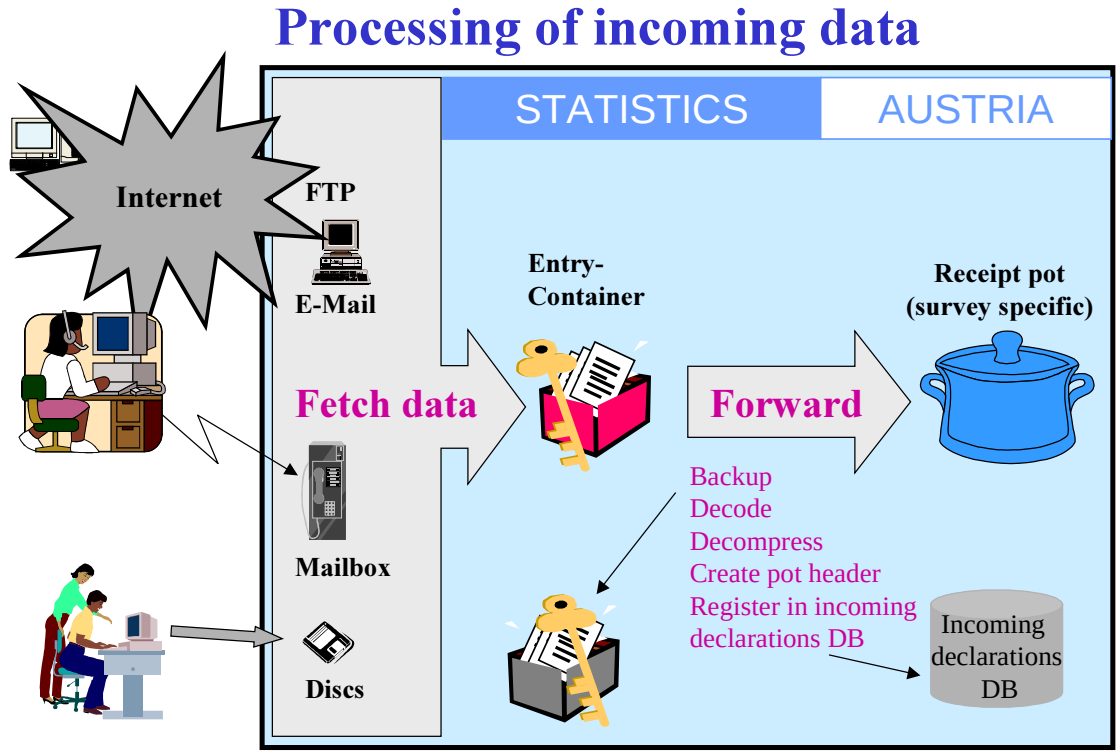

Fig. 2

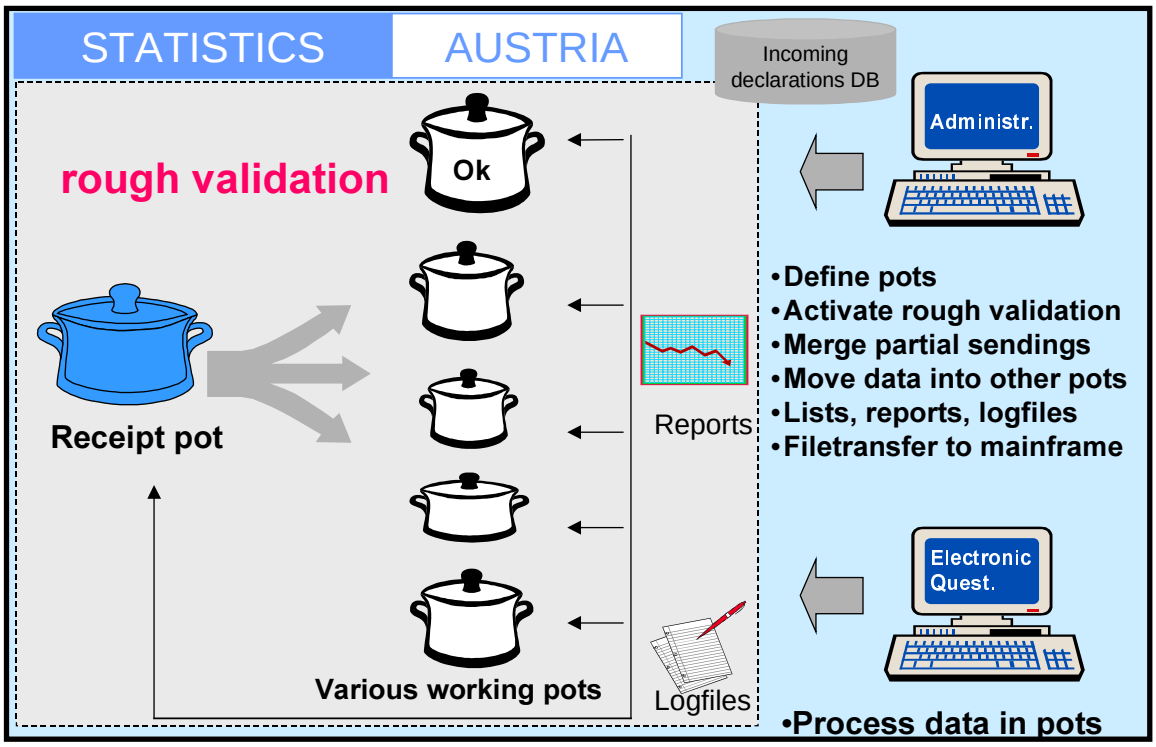

Fig. 3

As a number of functions (e.g. compressing/decompressing, encoding/decoding of data) will be required by EFBMS as well as by other SDSE programs – and probably also by software which will be developed at Statistics Austria in the future – these functions will be realized as reusable components.

III. General requirements

Here are some of the requirements that must be met by EFBMS.

1. EFBMS must be usable for diverse statistical surveys of different degrees of complexity (including the highly complex economic surveys *structural business statistics* and *short term statistics*). If a respondent is obliged to report for several surveys, he/she must not be compelled to install EFBMS more than once, but an EFBMS installation should enable the collection and administration of the response data of different surveys. If there is a new survey, only the registration of the metadata describing this survey and possibly some specific components (e.g., search of a classification code) should be necessary.

2. Sometimes a respondent entrusts another person or company (a third party declarant, for example, an accountant firm) to fill in the questionnaires and send them to Statistics Austria. As a third party declarant may be active for more than one client, EFBMS must enable the collection and administration of data for several respondents.

3. EFBMS must offer a local and a network installation variant. In local installation, the user interface program and the database are located on one PC. In network installation, the data – which comprise statistical response data as well as all

metadata – are stored on a server accessible to a number of users working on different PCs. EFBMS must guarantee that a questionnaire which is edited by one user is locked against write access of others.

4. For the storage of data, a relational database management system will be part of EFBMS (this will be the Microsoft Database Engine MSDE, a simplified version of SQL Server which can be deployed free of charge). But it will also be possible to use an existing database server like Oracle, DB2 or SQL Server instead of MSDE.

5. If statistical response data are confidential within the company of the respondent, it must be possible to define quite sophisticated access rights. But to keep the program simple (especially for small and medium enterprises), the user management and authorization features of EFBMS will not be activated by default.

6. A statistical survey can consist of more than one questionnaire. For example, in the *structural business statistics* survey there are three types of observation units (enterprise, establishment and local unit of employment), and for each observation unit a questionnaire has to be answered. Moreover, different kinds of relationships exist between these units: for example, some enterprises consist of several establishments consisting of several local units of employment, whereas other enterprises do not have an establishment but only local units of employment, and so on. EFBMS must support these hierarchical relations between observation units.

7. With some statistical surveys, it is just a matter of distributing empty questionnaires to the respondents, and the respondents decide for themselves which and how many of them they must fill in. With regard to business statistics, however, it is within the responsibility of Statistics Austria to determine which establishments and which local units of employment an enterprise consists of. According to this given structure, the respondent currently receives a corresponding number of paper questionnaires containing pre-printed data (e.g. identification code of the observation unit, address, NACE code, etc.). If a survey is carried out electronically, the same initialization must be possible with the structure of the respondent and with respondent-specific data. EFBMS must guarantee that, where a survey with obligatory initialization is concerned, a respondent is able to generate the questionnaires of a survey period only if these respondent-specific data can be provided. These data (which will be encoded by a symmetric encryption algorithm) will be sent out by Statistics Austria via e-mail or distributed on a CD-ROM. Later the respondents will also have the possibility to download them from the world wide web.

8. To keep EFBMS flexible and expansible, it will be realized in component architecture utilizing Microsoft's *Component Object Model* (COM).

9. All questionnaires of a survey – including validity checks and actions triggered by certain events (for example, the automatic calculation of the sum of numerical values entered by the user, or changes in attributes of questions like visible/invisible or enabled/disabled) – will be defined in XML syntax. A special component of EFBMS is responsible for the interpretation of these parameter files and for the dynamic generation and presentation of actual questionnaire windows. Thus, when a new survey is prepared for electronic data collection, no program source code has to be written or changed. Expert statisticians will define the necessary metadata for

EFBMS without the help of IT staff members (as long as there are no new components needed – see next item).

10. With some surveys users must be able to search for classification codes (like NACE or PRODCOM). As classifications often are quite large (and can contain further metadata like extensive descriptions of the classification members or a list of terms connected to them), they will be distributed as COM components responsible for the presentation of the classification (including and offering different ways of searching for a specific item) and for checking the validity of a code entered by the user. These classification components are called upon by EFBMS and communicate with EFBMS via pre-set interfaces; as long as the interface methods are the same it will be possible to deploy new classification components without the need of changing EFBMS source code.

11. As classifications may change in the course of time, the mentioned classification components must administrate several versions of a classification. An already installed component must be open to take up the data of a new classification version later on.

12. Automatic completion of the questionnaires must be a primary goal, in particular with extensive surveys which take place periodically. For this purpose, the respondent must be permitted to supply the response data via his/her own EDP system. The data must be provided in the standardized EFBMS import/export format which – like the response format used for transmitting the data to Statistics Austria – will be defined in XML syntax.

13. With regard to data validation, in case of a survey with hierarchically related observation units it must be possible to define validation rules across those hierarchical levels (e.g., the number of employees in an enterprise questionnaire must be equal to the sum of the numbers of employees in the establishment questionnaires).

14. There must be two types of validation rules: those which force the users to correct any errors found, and those which enable the respondents to insist on their answers, although the data conflict with a rule. In the latter case the respondent will have the opportunity to attach a note explaining why he/she thinks that the answers are correct.

15. The respondent must be able to print questionnaires, but these printouts are only for internal use and will not be accepted by Statistics Austria.

16. When a respondent wants to send his/her response data to Statistics Austria (by e-mail, FTP or dial line connection), EFBMS automatically performs the defined validation checks if the user has not yet activated them manually. Then the XML message is generated, compressed and encoded by an asymmetric encryption algorithm. To control the correct data transfer, a control value will be computed and added to the transmission data. After sending the data, the respondent will receive a transmission receipt.

17. EFBMS will run on the 32 bit Windows platform (Windows 95, Windows 98, Windows NT 4, Windows 2000).

~~IV.~~ Deployment

The SDSE will be first used in January 2001 for *short term statistics*, a monthly survey with almost 20 000 respondents. Together with the paper questionnaires, every respondent will receive a CD-ROM (containing EFBMS, two classification components for PRODCOM and NACE, structure data defining the survey and its questionnaires, and encrypted respondent-specific initialization data) and a code which is necessary to access the initialization data.

After the installation of the program and the loading of the *short term statistics* metadata and the initialization data the software is ready for use.

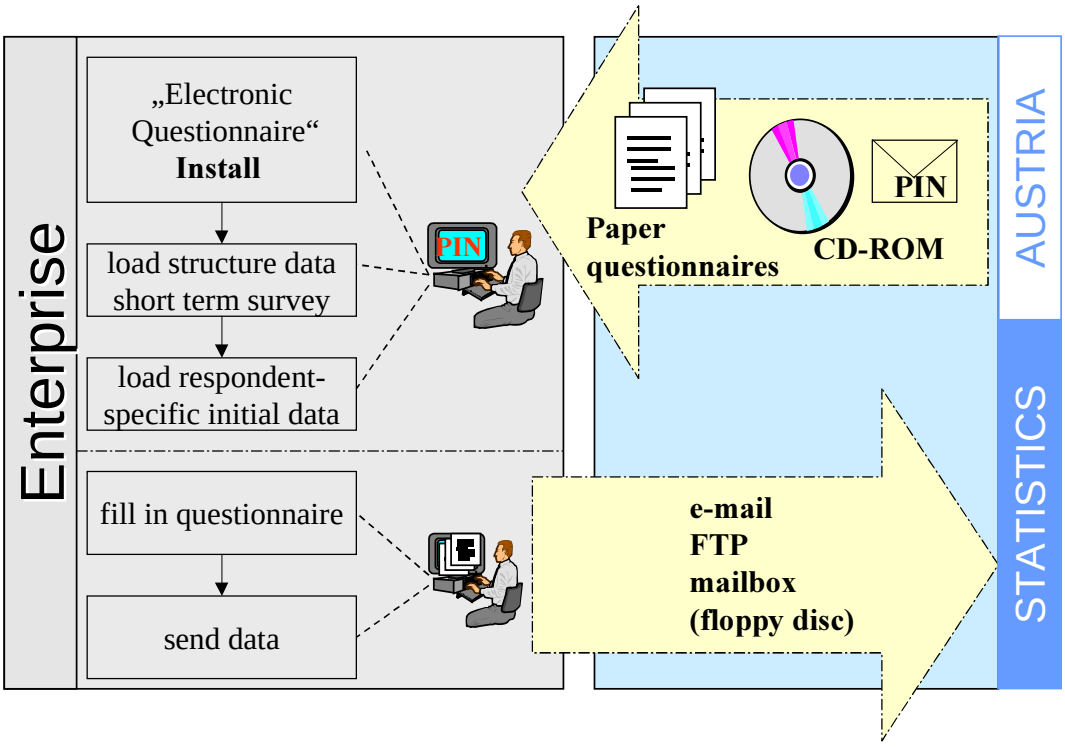

Fig. 4

~~V.~~IV. Current state

It is planned that CSC Austria will complete the beta version of EFBMS at the end of September so that Statistics Austria will be able to perform internal tests. Because of the very tight schedule analysis of a designing tool for questionnaires – which should have started by now – had to be postponed. As XML metadata may be edited by a text editor, as well, this postponement does not change our deployment schedule.

The following screenshots are from September 8 and show the alpha version of EFBMS. In Fig. 5 the main window of EFBMS can be seen. The navigation window on the left displays respondents, surveys, questionnaires and so on in a tree view. For every node in this tree, a click with the right mouse button opens a popup menu. The info window on the bottom displays the context help of a question as well as a list of

validation errors (where the users can insist on their answers), and it enables the user to attach notes to questions. Figures 6 – 7 show some pages of the dynamically generated questionnaire for an observation unit of type “enterprise“. Fig. 8 presents the PRODCOM classification component, where the classification is displayed in a tree view. There are several alternatives to search for a code (text search, search in a list of related terms, search by a code of the combined nomenclature).

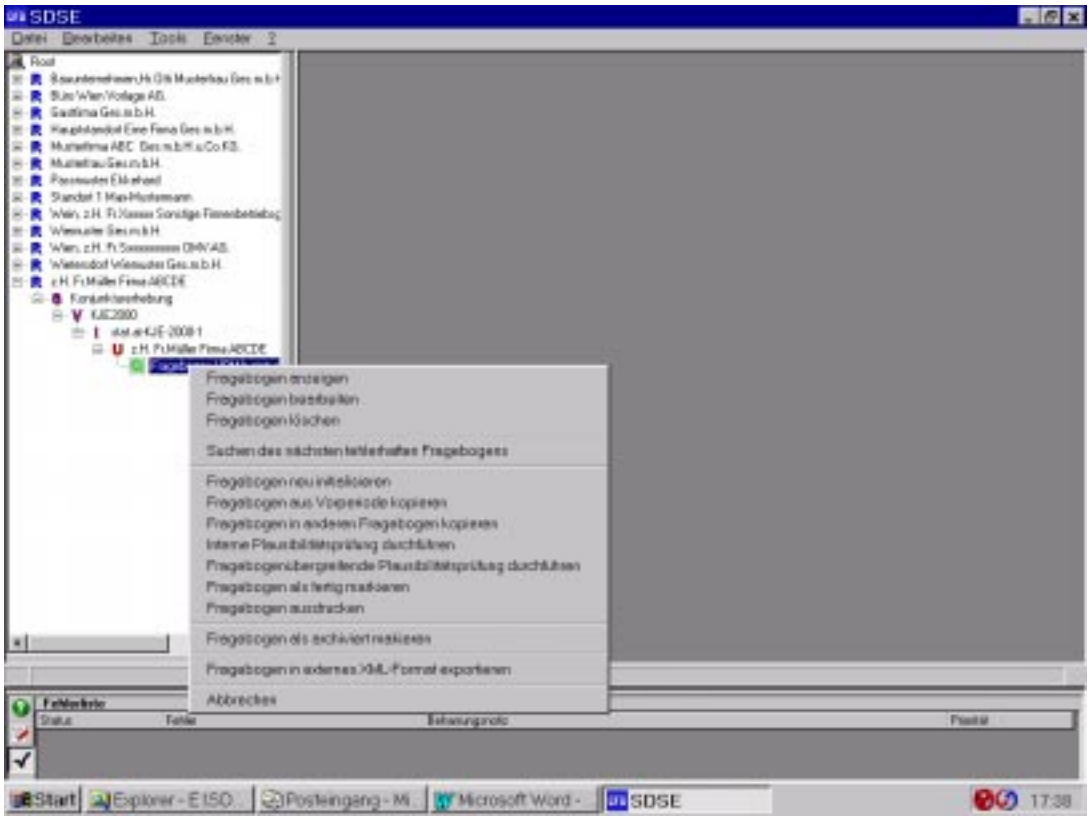

Fig. 5

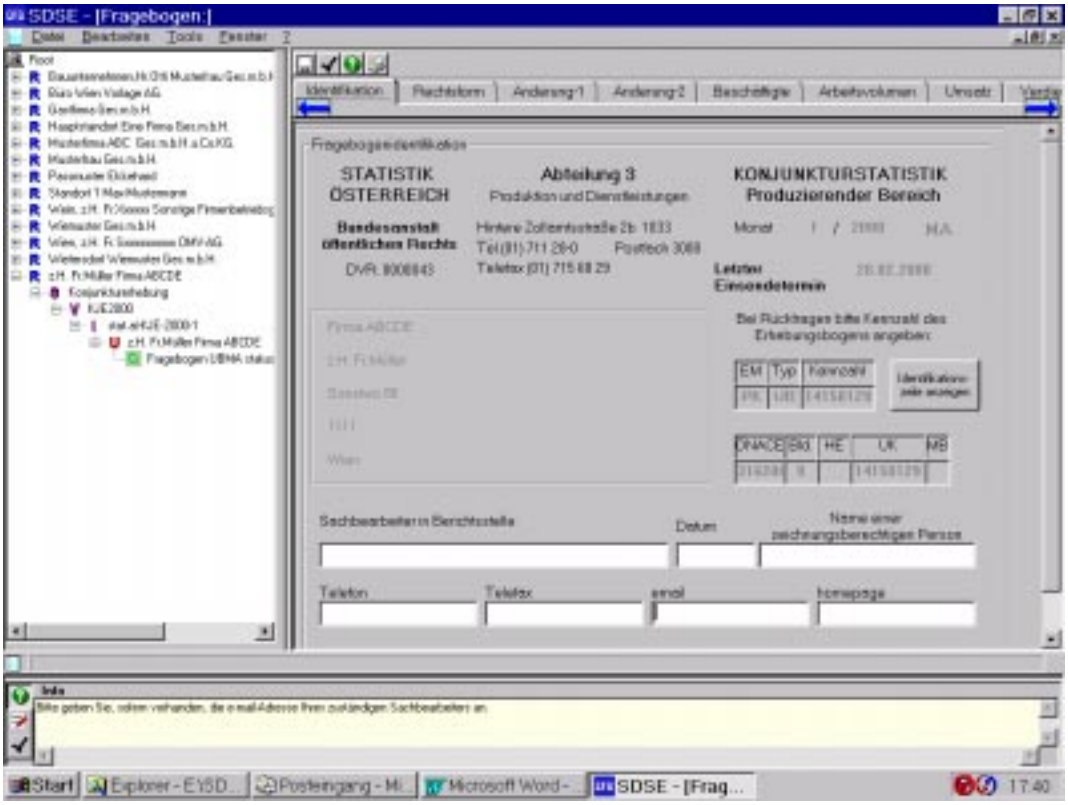
SDSE - [Fragebogen:]
STATISTIK ÖSTERREICH
Abteilung 3
Produktion und Dienstleistungen
KONJUNKTURSTATISTIK
Produzierender Bereich
Sachbearbeiter/in Berichtsstelle
Datum
Telefon
Telefax
email
Homepage

Fig. 6

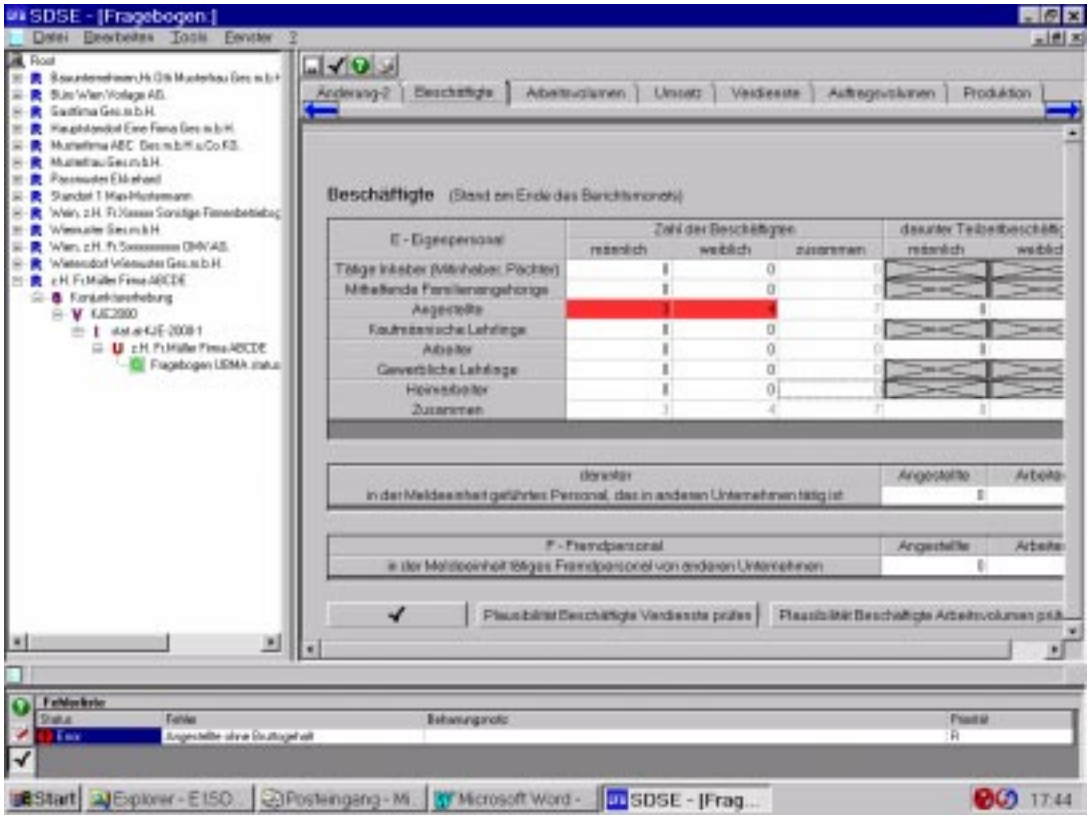
SDSE - [Fragebogen:]
Beschäftigte (Stand am Ende des Berichtsmonats)
E - Eigenpersonal
Zahl der Beschäftigten
männlich
weiblich
zusammen
Tätige Inhaber (Mitinhaber, Pächter)
Mithelfende Familienangehörige
Angestellte
Kaufmännische Lehrlinge
Arbeiter
Gewerbliche Lehrlinge
Heimarbeiter
Zusammen
F - Fremdpersonal

Fig. 7

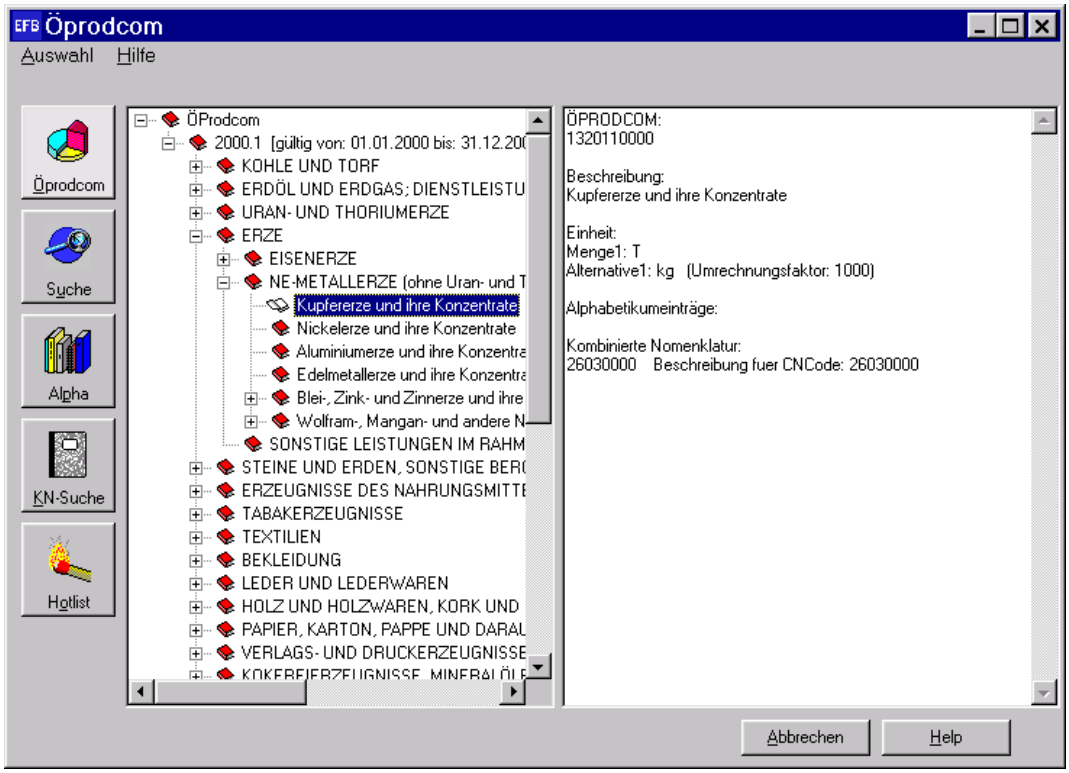

Fig. 8

Vienna, 14 September 2000