SDSE

(Extract from CoRD 026, 'Developing a system for electronic raw data collection at Statistics Austria', W Koller and G Zettl, 2000)

1. System overview

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

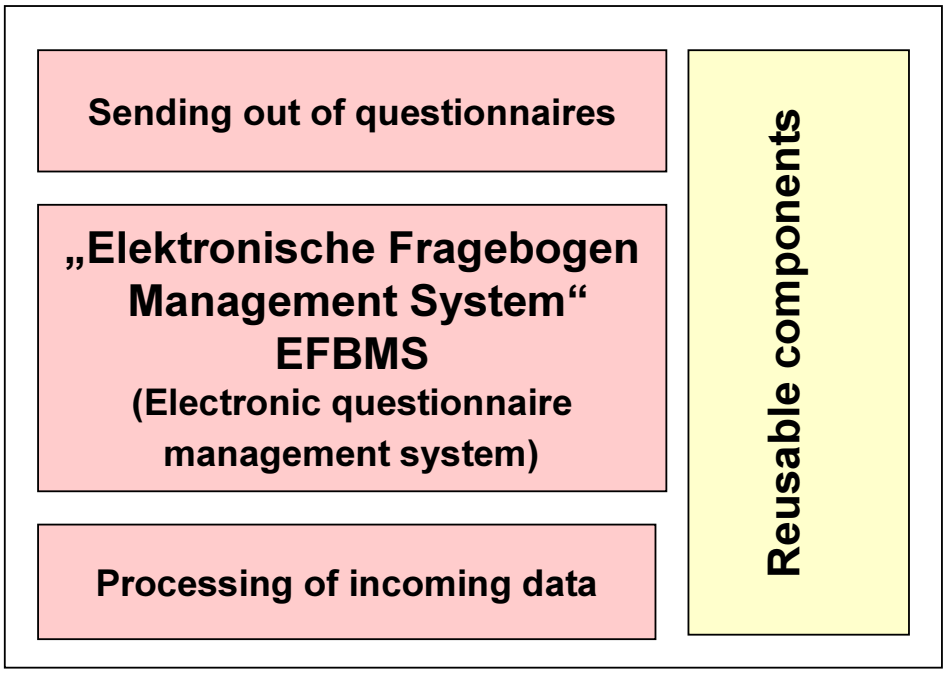

Fig. 1

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.

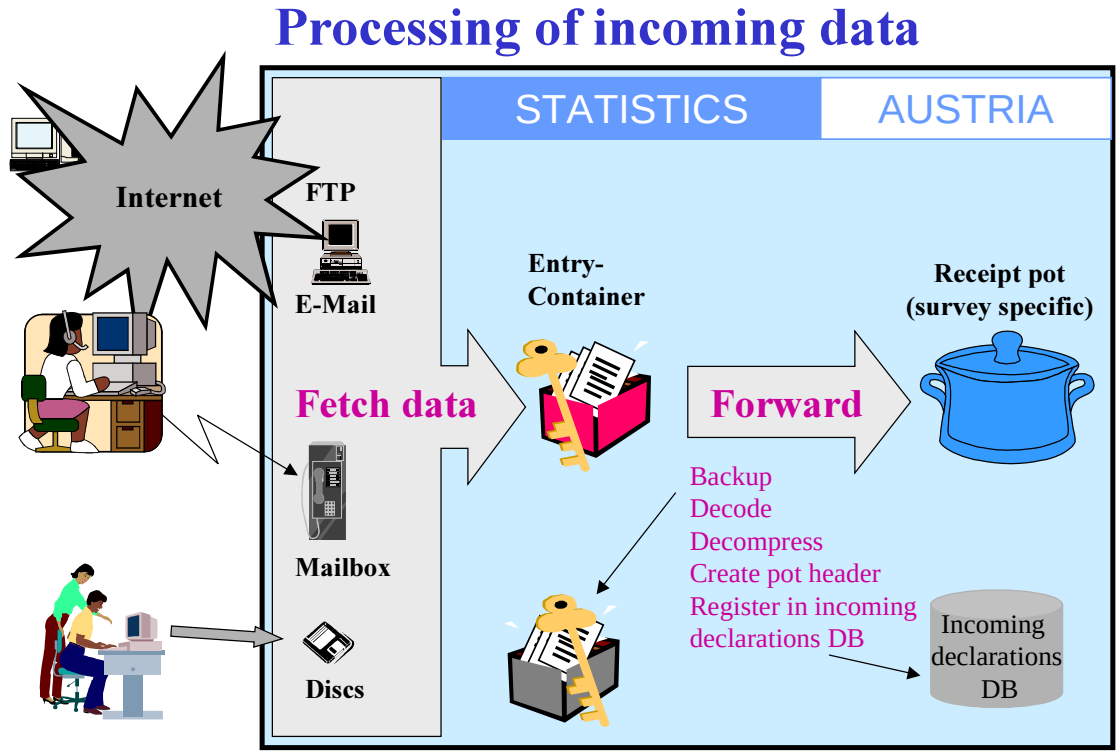

Fig. 2

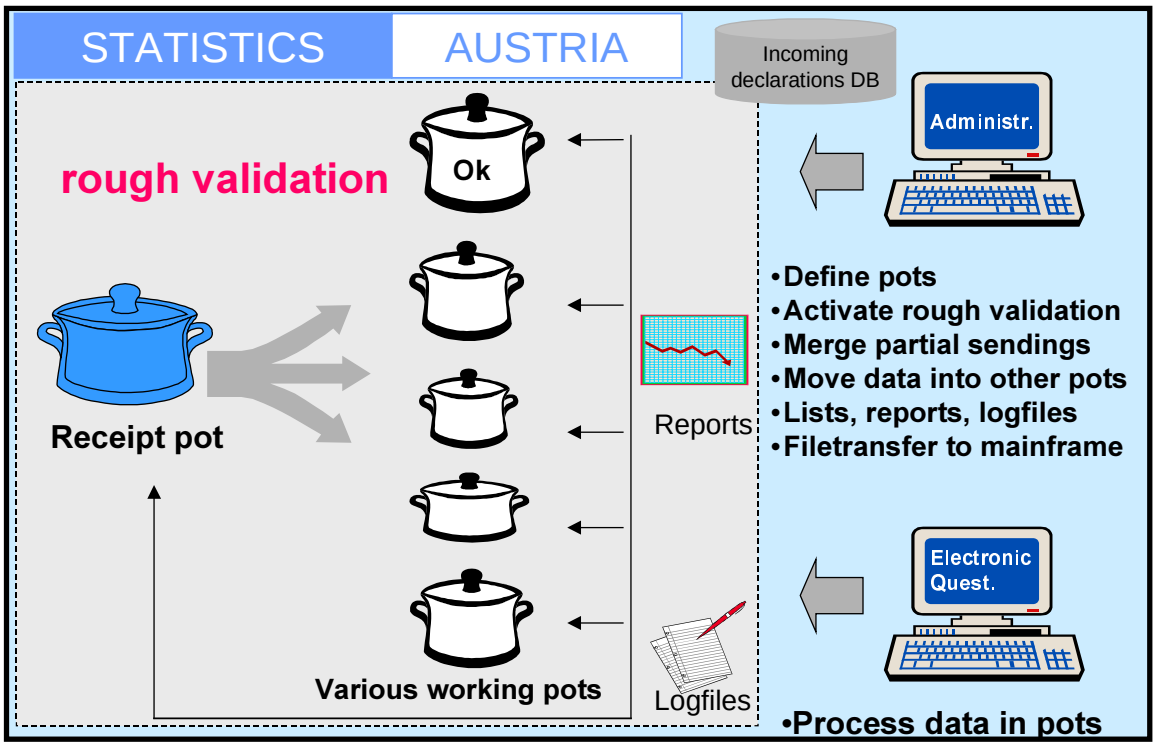

Fig. 3

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.

2. General requirements

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

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.

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.

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.

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.

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.

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.

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.

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

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).

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.

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.

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.

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).

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.

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

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.

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

3. 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.

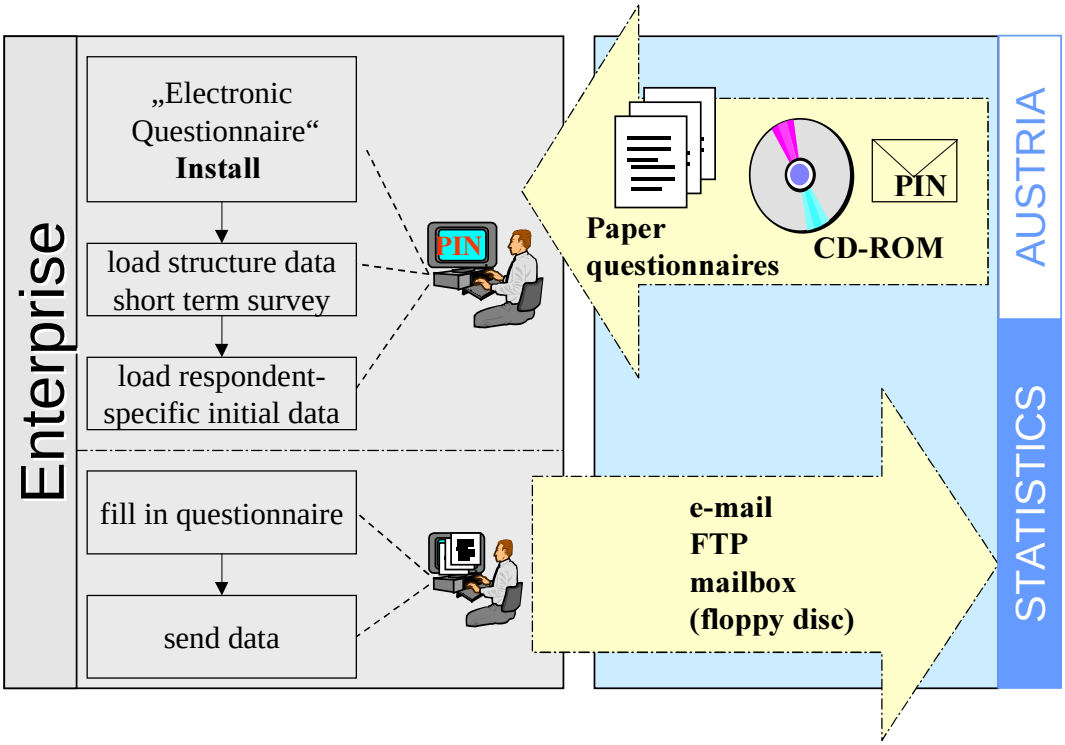
Enterprise
„Electronic Questionnaire“ Install
load structure data short term survey
load respondent-specific initial data
fill in questionnaire
send data
PIN
Paper questionnaires
CD-ROM
PIN
e-mail
FTP
mailbox
(floppy disc)
AUSTRIA
STATISTICS

Fig. 4