STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 23
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (ii): Metadata modelling and terminology issues

DEVELOPING A METADATA-BASED SYSTEM FOR ELECTRONIC RAW DATA COLLECTION AT STATISTICS AUSTRIA

Submitted by Statistics Austria [1]

Contributed paper

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

I. HISTORY

1. In 1998, Statistics Austria was experiencing a lot of political pressure (for example, from the Chamber of Commerce and Industry and from the Ministry of Economic Affairs) to provide software for collecting and transmitting statistical data to the respondents of business surveys free of charge in order to relieve the enterprises. Later on, this request became part of the "Federal Statistics Act 2000".[2]

2. In June 1998, a working group was founded consisting of members of the IT division and of the division 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 structural business statistics, an annual survey with approximately 40 000 respondents.

3. Statistics Austria had neither the manpower nor the PC programming skills to develop the software in its IT division, so we therefore began to prepare a call for tenders. In the course of this work the scope of the project grew 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.

---

[1] Prepared by Wolfgang Koller and Günther Zettl.

[2] I. § 28 (3): „Auf Wunsch sind den Auskunftspflichtigen die entsprechenden Unterlagen für die Auskunfts-erteilung 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").

4. 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-related information including questionnaires and validity checks in XML parameter files. Respondents and intermediaries (third party declarants) such as accountant firms will receive 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. There are plans for the program to be used by Statistics Austria as well, so that our statisticians will be able to view, check and edit incoming data with the same tool that the respondents are using.

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

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

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

8. The second phase of the call for tenders ended on October 18, 1999. Four of the six participating companies submitted concepts for the realisation 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.

9. 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 initial tests. This prototype included an alpha version of the PRODCOM classification component.

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

11. 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 prepared several marketing activities, and sent a letter with information about the project to every respondent of the short-term survey at the end of August, asking respondents whether they intended to use the program. The results of this survey are presented in chapter VI. We also had a first meeting on the subject of supporting the import interface of our questionnaire software in SAP R/3.

12. In cooperation with an advertising agency, some promotional material was designed (e.g. the label printed on the CD-ROM and a public relations folder) in September and October 2000. Since "e-whatever" is very fashionable today (e-Business, e-Commerce, e-Government, etc.), we decided to climb on the bandwagon and call the electronic questionnaire software e-Collect.

13. Meanwhile, CSC Austria continued the development of the software. A second prototype was finished in mid-August and the so-called alpha version (with many of the features of the final product

implemented) was deployed on September 5. Compared to the original schedule, software development is about three months behind because of the complexity of the system. This delay shortens the time available for tests and corrections. Nevertheless the project team members at Statistics Austria and CSC Austria are working very hard – at some times literally day and night – to meet the deadline (end of October) for the pilot test.

II. SYSTEM OVERVIEW

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

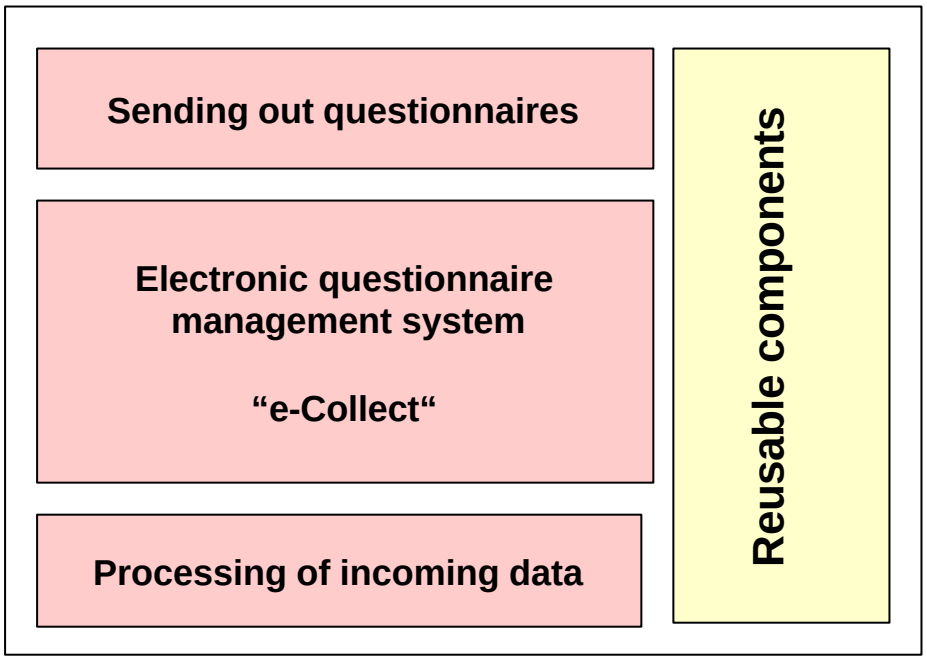

Fig. 1

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

- structural metadata (describing the collector of statistical information, the survey and its various versions, the types of observation units and their respective questionnaire types (including validity checks), hierarchical relationships between observation unit types, and some other objects);
- respondent-specific metadata (describing the respondent, the actual observation units for which the respondent must complete questionnaires, the actual relationships between them, and so on). These files can also contain statistical data imported automatically into new questionnaires – the so-called initialisation data.

Another part of this sub-system will be a tool for the design of questionnaires and for the management of structural XML metadata ("e-Collect metadata management system").

16. The e-Collect program is the most important – and most complex – component of the SDSE. It will be provided to 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 (in the future, it will probably be made available to other institutions as well using e-Collect for their own

surveys). The staff of Statistics Austria should also be able to use it for viewing and processing of the transferred data.

17. The third sub-system "Processing of incoming data" consists of programs which carry the statistical declarations from e-mail, FTP and mailbox servers at 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 organisational unit (fig. 2). Statisticians will have an online application to administer the incoming response data files (tentatively called the "pot application" according to pot cliparts which we included in Powerpoint slides to symbolize containers for response data). e-Collect will be used to view and to correct the contents of a file (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.

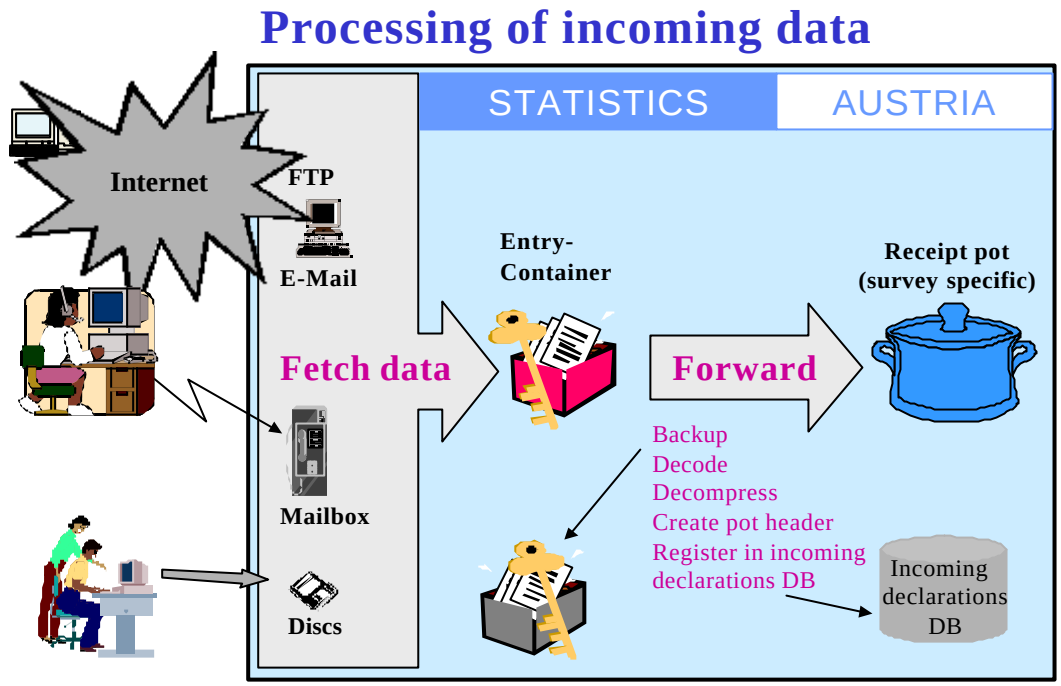
Processing of incoming data
STATISTICS
AUSTRIA
Internet
FTP
E-Mail
Mailbox
Discs
Entry-
Container
Fetch data
Forward
Receipt pot
(survey specific)
Backup
Decode
Decompress
Create pot header
Register in incoming
declarations DB
Incoming
declarations
DB

Fig. 2

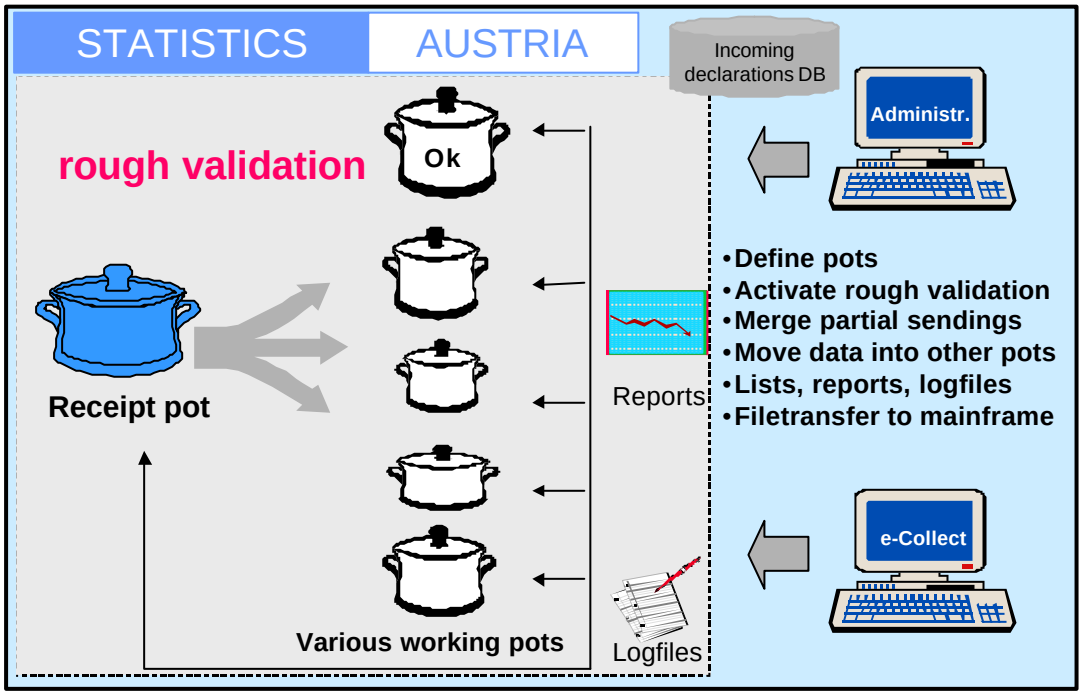
STATISTICS
AUSTRIA
Incoming
declarations DB
Administr.
rough validation
Ok
Receipt pot
Reports
•Define pots
•Activate rough validation
•Merge partial sendings
•Move data into other pots
•Lists, reports, logfiles
•Filetransfer to mainframe
e-Collect
Various working pots
Logfiles

Fig. 3

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

III. GENERAL REQUIREMENTS

19. The following paragraphs in this section list some of the requirements of e-Collect.

20. e-Collect must be usable for diverse statistical surveys of different degrees of complexity (including the very complex economical surveys of structural business statistics and short-term statistics). If a respondent is obliged to report for several surveys, one installation of e-Collect should suffice. For any new survey, the registration of the metadata describing this survey and, possibly, the installation of some specific components (e.g., for searching a classification code) should be enough.

21. Sometimes a respondent delegates another person or company (a third party declarant; for example, an accountant firm) to complete the questionnaires and send them to Statistics Austria. Since a third party declarant may be active for more than one client, e-Collect must enable the collection and administration of data for several respondents.

22. e-Collect must offer a local version as well as a network installation variant. With local installation, the user interface program and the database are located on one PC. Network installation stores the data – which comprise statistical response data as well as all metadata – on a server accessible to a number of users working on different PCs. e-Collect must guarantee that a questionnaire which is being edited by one user is protected against write access by other users.

23. A relational database management system will be part of e-Collect for the storage of data (this system will be the Microsoft Database Engine MSDE, a simplified version of SQL Server which can be deployed free of charge). Nevertheless, it will be possible to use an existing database server as well, such as Oracle, DB2 or SQL Server instead of MSDE.

24. If statistical response data are confidential within the company of the respondent, it must be possible to define relatively sophisticated access rights. To keep the program simple (especially for small- and medium-sized enterprises) however, the user management and authorization features of e-Collect will not be activated by default.

25. A statistical survey may consist of more than one questionnaire. For example, the structural business survey comprises three types of observation units (enterprise, establishment and local unit of employment), and for each observation unit a questionnaire has to be completed. Moreover, different kinds of relationships exist between these units. Some enterprises consist of several establishments comprising various local units of employment, whereas other enterprises have no establishment at all except local units of employment, and so on. e-Collect must support these hierarchical relations between observation units.

26. 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 the responsibility of Statistics Austria to determine which establishments and which local units of employment comprise an enterprise. 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. e-Collect must guarantee that, where a survey with obligatory initialisation is concerned, a respondent will be 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 CD-ROM. Later on, the respondents will be able to download them from the world wide web.

27. To keep e-Collect flexible and capable of being extended, it is produced in component architecture utilising Microsoft's Component Object Model (COM).

28. 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 e-Collect is responsible for the interpretation of this information 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. Statisticians will define the metadata necessary for e-Collect with little or no help from IT staff members – as long as there are no new components required.

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

30. As classifications may change over the course of time, the classification components mentioned here must administer to several versions of a classification. An already installed component must be open to data from a new classification version later on.

31. 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 a standardised e-Collect import/export format which – like the response format used for transmitting the data to Statistics Austria – is defined in XML syntax.

32. As far as the data validation of a survey with hierarchically related observation units is concerned, validation rules across those hierarchical levels must be definable (e.g. the number of employees in an enterprise questionnaire must be equal to the sum of employees in the establishment questionnaires).

33. There will be three types of validation rules: those which warn the user of a possible error, those which force the user to correct any errors found, and those which enable the respondent to insist on his/her answer, 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 answer is correct.

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

35. When a respondent wants to send his/her response data to Statistics Austria (by e-mail, FTP or dial line connection), e-Collect ensures that the questionnaires have been validated and the user has corrected all errors (or insisted on his/her answers). 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 out, the respondent will receive a transmission receipt.

36. When e-Collect is used by Statistics Austria to view and to correct response data, the original value entered by the respondent must not be lost whenever an answer is changed by a statistician. This means that e-Collect must be able to store a history of changes for each data item. Statisticians must also be able to attach notes to certain questions.

37. e-Collect will run on the 32 bit Windows platform (Windows 95, Windows 98, Windows NT 4, Windows 2000).

IV. DEPLOYMENT

38. SDSE is scheduled for its first use in January 2001, in short-term statistics for a monthly survey with almost 20 000 respondents. Together with the paper questionnaires, every respondent will receive a CD-ROM (containing e-Collect, two classification components for PRODCOM and NACE, metadata defining the survey and its questionnaires, and encrypted respondent-specific initialisation data) and a code necessary to access the initialisation data.

39. The CD-ROM will also comprise an "auto-run" program developed by Statistics Austria, offering the users several options; for example, they may install e-Collect and other free software like Adobe Acrobat Reader, or they may view the installation manual and several "quick start" tutorials. We also plan to include further information on the CD-ROM accessible from this auto-run program (e.g. information about Statistics Austria, results of previously conducted surveys, and so on – possibilities are restricted only by the storage space available on the disc).

40. After the installation of the program and the loading of short-term statistics metadata and initialisation data, the software is ready for use.

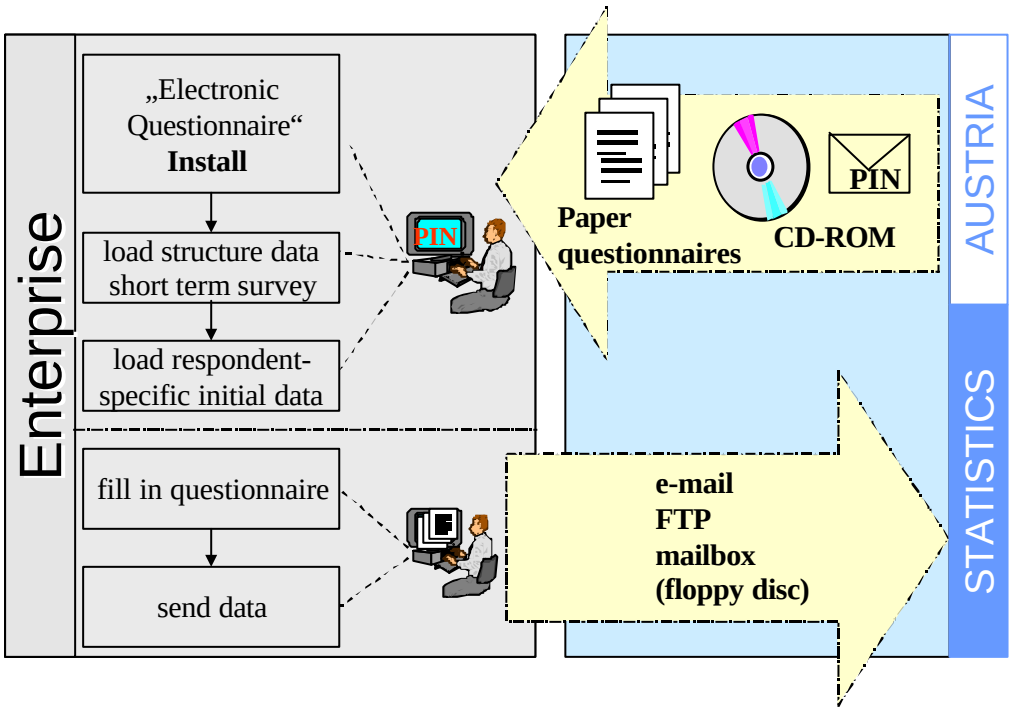

Fig. 4

V. CURRENT STATE

41. The SDSE project always had a very tight schedule, with most tasks situated on the critical path. Because some activities took more time than expected, non-essential working steps such as analysis and development of a design tool for questionnaires (XML metadata can be edited by a simple text editor as well) had to be postponed.

42. Unfortunately, in October, Murphy's Law struck, and some newly discovered bugs prevented us from finishing the pilot test version of e-Collect in time. Now the pilot test is scheduled to start in mid-November. We will have to wait for the results before taking a decision as to whether the final deployment will have to be delayed too.

43. The following screenshots are from October 30 and show the beta version of e-Collect. Fig. 5 presents the main window of e-Collect. The navigation window on the left displays respondents, surveys, questionnaires and so on in tree form. For every node in this tree, a click with the right mouse button opens a pop-up menu. The info window on the bottom displays the summarised help to a question, as well as a list of validation errors (where the users can insist on their answers), and it enables the users to attach notes to questions. Figures 6 and 7 show some pages of the dynamically generated questionnaire for an observation unit of the "enterprise" type. Fig. 8 presents the PRODCOM classification component, displaying the classification again in tree form. There are several alternatives to search for a code (search in a list of more than 30 000 synonyms, text search, or search by a code of the Combined Nomenclature). Fig. 9 is a screenshot of one of the "quick start" tutorials accessible from the auto-run program.

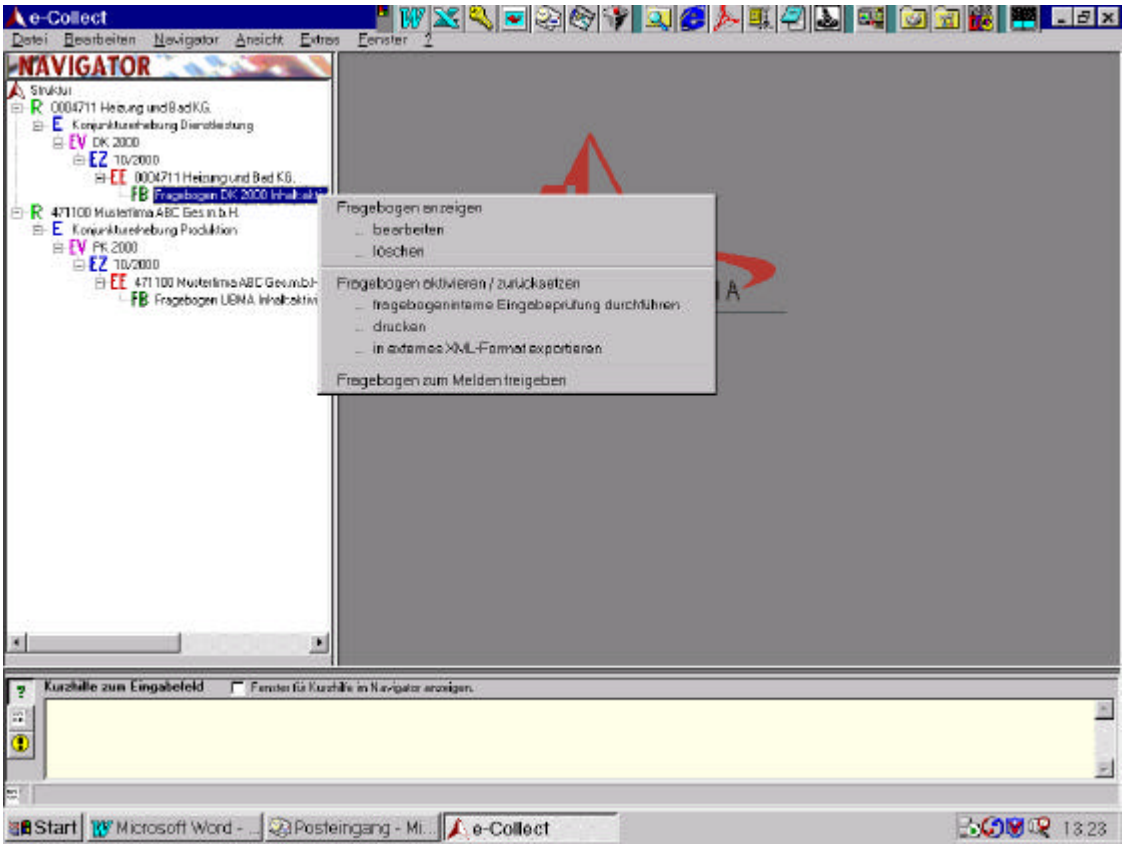

Fig. 5

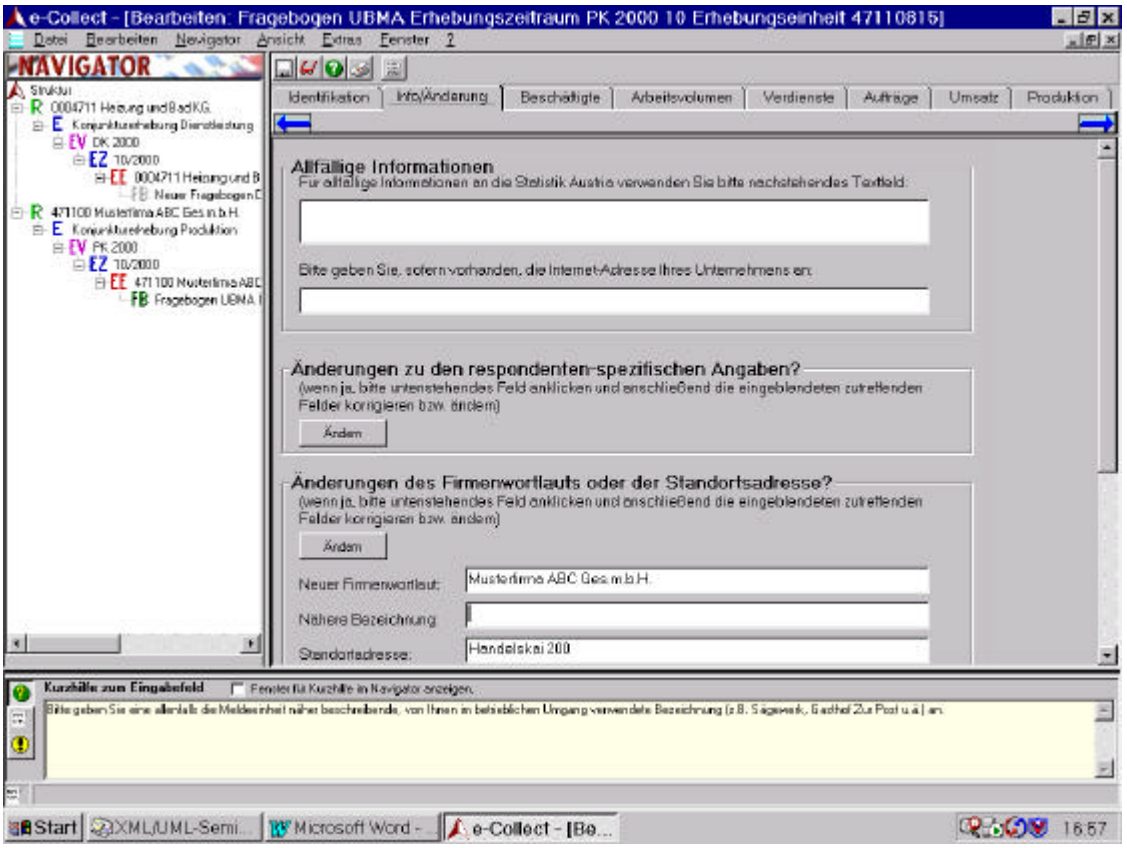
e-Collect - [Bearbeiten: Fragebogen UBMA Erhebungszeitraum PK 2000 10 Erhebungseinheit 47110815]
Datei Bearbeiten Navigator Ansicht Extras Fenster ?
NAVIGATOR
Identifikation
Info/Änderung
Beschäftigte
Arbeitsvolumen
Verdienste
Aufträge
Umsatz
Produktion
Allfällige Informationen
Für allfällige Informationen an die Statistik Austria verwenden Sie bitte nachstehendes Textfeld:
Bitte geben Sie, sofern vorhanden, die Internet-Adresse Ihres Unternehmens an:
Änderungen zu den respondenten-spezifischen Angaben?
(wenn ja, bitte untenstehendes Feld anklicken und anschließend die eingeblendeten zutreffenden Felder korrigieren bzw. ändern)
Ändern
Änderungen des Firmenwortlauts oder der Standortsadresse?
(wenn ja, bitte untenstehendes Feld anklicken und anschließend die eingeblendeten zutreffenden Felder korrigieren bzw. ändern)
Ändern
Neuer Firmenwortlaut:
Musterfirma ABC Ges.m.b.H.
Nähere Bezeichnung:
Standortadresse:
Handelskai 200
Kurzhilfe zum Eingabefeld
Fenster für Kurzhilfe im Navigator anzeigen.
Start
XML/UML-Semi...
Microsoft Word -...
e-Collect - [Be...
16:57

Fig. 6

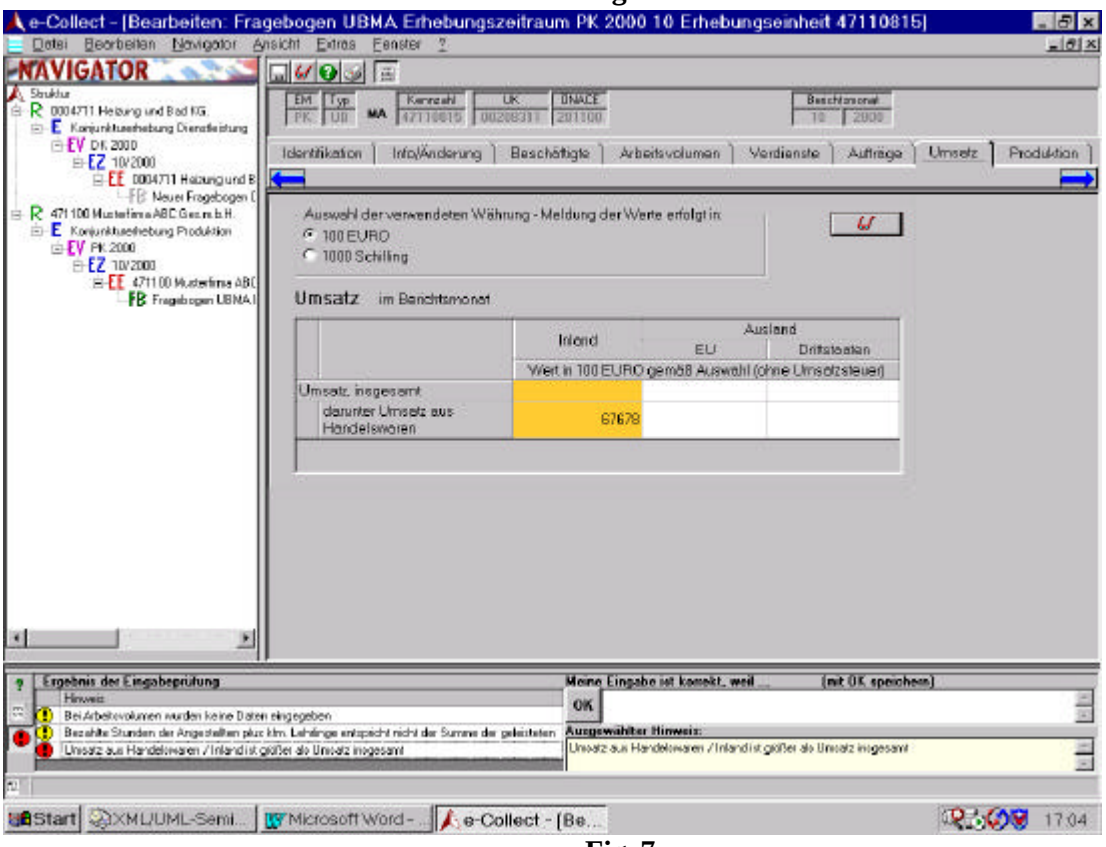
e-Collect - [Bearbeiten: Fragebogen UBMA Erhebungszeitraum PK 2000 10 Erhebungseinheit 47110815]
Datei Bearbeiten Navigator Ansicht Extras Fenster ?
NAVIGATOR
Identifikation
Info/Änderung
Beschäftigte
Arbeitsvolumen
Verdienste
Aufträge
Umsatz
Produktion
Auswahl der verwendeten Währung - Meldung der Werte erfolgt in:
100 EURO
1000 Schilling
Umsatz im Berichtsmonat
Inland
Ausland
EU
Drittstaaten
Wert in 100 EURO gemäß Auswahl (ohne Umsatzsteuer)
Umsatz, insgesamt
darunter Umsatz aus Handelswaren
67679
Ergebnis der Eingabeprüfung
Hinweis
Bei Arbeitsvolumen wurden keine Daten eingegeben
Bezahlte Stunden der Angestellten plus Ihm. Lehrlinge entspricht nicht der Summe der geleisteten
Umsatz aus Handelswaren / Inland ist größer als Umsatz insgesamt
Meine Eingabe ist korrekt, weil ... (mit OK speichern)
OK
Ausgewählter Hinweis:
Umsatz aus Handelswaren / Inland ist größer als Umsatz insgesamt
Start
XML/UML-Semi...
Microsoft Word -...
e-Collect - [Be...
17:04

Fig. 7

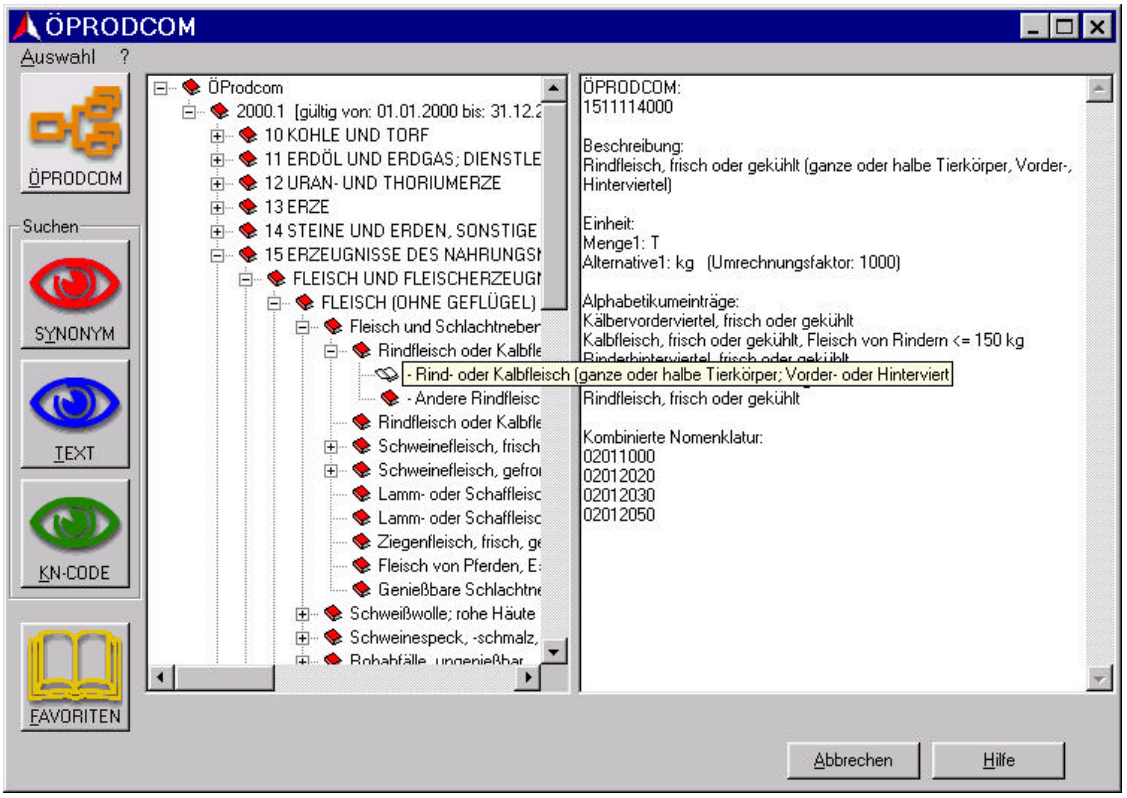
ÖPRODCOM
Auswahl ?
ÖPRODCOM
Suchen
SYNONYM
TEXT
KN-CODE
FAVORITEN
ÖProdcom
2000.1 [gültig von: 01.01.2000 bis: 31.12.2
10 KOHLE UND TORF
11 ERDÖL UND ERDGAS; DIENSTLE
12 URAN- UND THORIUMERZE
13 ERZE
14 STEINE UND ERDEN, SONSTIGE
15 ERZEUGNISSE DES NAHRUNGSM
FLEISCH UND FLEISCHERZEUGI
FLEISCH (OHNE GEFLÜGEL)
Fleisch und Schlachtneber
Rindfleisch oder Kalbfle
- Rind- oder Kalbfleisch (ganze oder halbe Tierkörper; Vorder- oder Hinterviert
- Andere Rindfleisc
Rindfleisch oder Kalbfle
Schweinefleisch, frisch
Schweinefleisch, gefro
Lamm- oder Schaffleisc
Lamm- oder Schaffleisc
Ziegenfleisch, frisch, ge
Fleisch von Pferden, E
Genießbare Schlachtne
Schweißwolle; rohe Häute
Schweinespeck, -schmalz,
ÖPRODCOM:
1511114000
Beschreibung:
Rindfleisch, frisch oder gekühlt (ganze oder halbe Tierkörper, Vorder-, Hintervierte
Einheit:
Menge1: T
Alternative1: kg (Umrechnungsfaktor: 1000)
Alphabetikumeinträge:
Kälbervorderviertel, frisch oder gekühlt
Kalbfleisch, frisch oder gekühlt, Fleisch von Rindern <= 150 kg
Rindfleisch, frisch oder gekühlt
Kombinierte Nomenklatur:
02011000
02012020
02012030
02012050
Abbrechen
Hilfe

Fig. 8

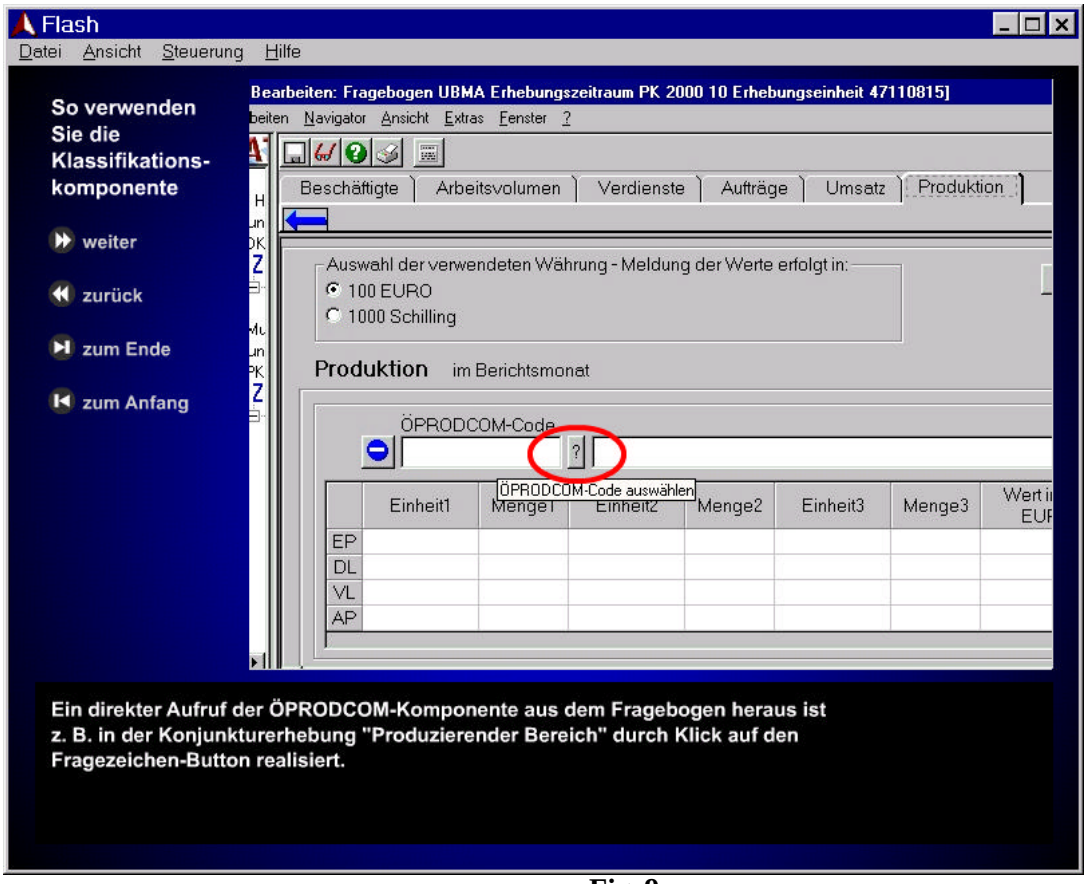

Fig. 9

VI. RESULTS OF THE SURVEY ON THE INTENDED USE OF E-COLLECT

44. As mentioned above, in August a short questionnaire was sent to 17 331 respondents of the short term survey. Within two months 7 506 questionnaires (among them 116 empty ones) were returned to Statistics Austria (response rate: 42.64 %).

45. Here are the results of this voluntary survey (because of multiple answers, sums may be higher than the number of returned questionnaires):

i) The monthly short-term statistics questionnaires are filled in

a) by one person in the company	5 623	76 %
b) by more than one person in the company	991	13 %
c) by a third party declarant (for example an accountant firm)	878	%

ii) Do you intend to use the electronic questionnaire software?

a) Yes, as soon as possible	2 899	39 %
b) Yes, but later	533	7 %
c) Yes, not for short-term statistics, but for		
c.1) structural business statistics	41	1 %
c.2) input statistics	10	0 %
c.3) other surveys	16	0 %
d) I have not made up my mind, yet	1 052	14 %
e) No, due to lack of necessary EDP equipment	1 029	14 %
f) No, I do not think it is useful for me	271	%

iii) Which form of data transmission to Statistics Austria will you use?

a) e-mail	3 205	43 %
b) FTP	113	2 %
c) Dial line connection (modem to modem)	716	10 %
d) floppy discs	329	

iv) On how many of your company's PCs will you use the electronic questionnaire software?

a) on one PC	3 066	41 %
b) on several PCs in a local area network	911	%

v) Statistics Austria intends to supply updates of the system via Internet (WWW). Would you take advantage of this possibility?

a) Yes	2 432	33 %
b) No, I would prefer a CD-ROM or floppy disc	1 169	6 %

vi) Do you use commercial software products in your company?

	1) SAP	2) Oracle	3) Baan	4) Peoplesoft	5) Other
A) Staff management	164	28	7	7	1 757
B) Payroll accounting	169	28	9	3	2 309
C) Stock management	233	41	29	3	1 727
D) Cost accounting	299	41	27		1 639
E) Invoicing	260	45	30	1	2 511
F) Profit and loss accounting	338	50	32	1	2 255

vii) Does your software support the EDIFACT format?

a) Yes	700	9 %
b) No	676	9 %
c) I do not know	2 419	33 %

VII. METADATA IN E-COLLECT

46. When discussing the concept of statistical metadata, the focus is often on the dissemination phase of the statistical life cycle. Metadata are primarily seen as data needed by end-users for the search, access and understanding of statistical information. A second main discussion topic is the creation and use of metadata within the organisations responsible for the production of statistics. In many surveys the source material and starting point for statistical information is data collected by means of questionnaires and, of course, the persons responsible for filling them in – the respondents – also require the metadata to support them in fulfilling their tasks. These metadata must be supplied by the survey designers.

47. In a complex electronic questionnaire program like e-Collect, there are several layers of metadata. A rough distinction can be made between metadata designed for respondents and metadata used by the software, although these two categories can overlap. For example: surveys and survey versions have attributes like ValidFrom, ValidTo, SurveyType (whether a survey is carried out periodically) and PeriodType (periodicity). These and other specifications are necessary to generate survey instances and to calculate the deadline for the responses to be sent to Statistics Austria, but they can also be presented to the users.

48. The first layer of metadata designed for respondents consists of information that is presented within the questionnaire form: the question text, answering hints, footnotes and so on. This layer exists in paper questionnaires as well as in electronic ones. In e-Collect, the information is defined within the XML description of the questionnaire (<QuestXML>) which itself is subdivided into five parts, namely definitions of:

a) questions and question groups (<QUDEF>);
b) the graphical user interface elements (<GUIDEF>);
c) the layout (<LAYOUTDEF>);
d) the event handling (<CTRLDEF>); and
e) the validation rules (<PLAUSDEF>).

49. Annex 1 presents an example of an excerpt from an XML questionnaire.

50. The second metadata layer is specific to e-Collect. When an edit field is focused, a short help text is displayed in the info window at the bottom of the screen. This text is defined in a file which is installed together with survey-specific help files.

51. The next layer consists of comprehensive explanations with regard to answering the questions, including definitions of terms and, sometimes, examples. For surveys carried out on paper, these descriptions are usually printed on separate pages. In e-Collect they are distributed as an extensive help system in HTML format. When the user presses the F1 button, the browser is launched and a help page relating to the active question is presented. In comparison with printed explanations, this help system offers all the advantages of a hypertext system. If Java is enabled in the browser, a hierarchically structured content frame is displayed in addition to the help pages, and the respondent can find a topic with the help of an index or by using full-text search. Fig. 10 shows a screenshot of the help system.

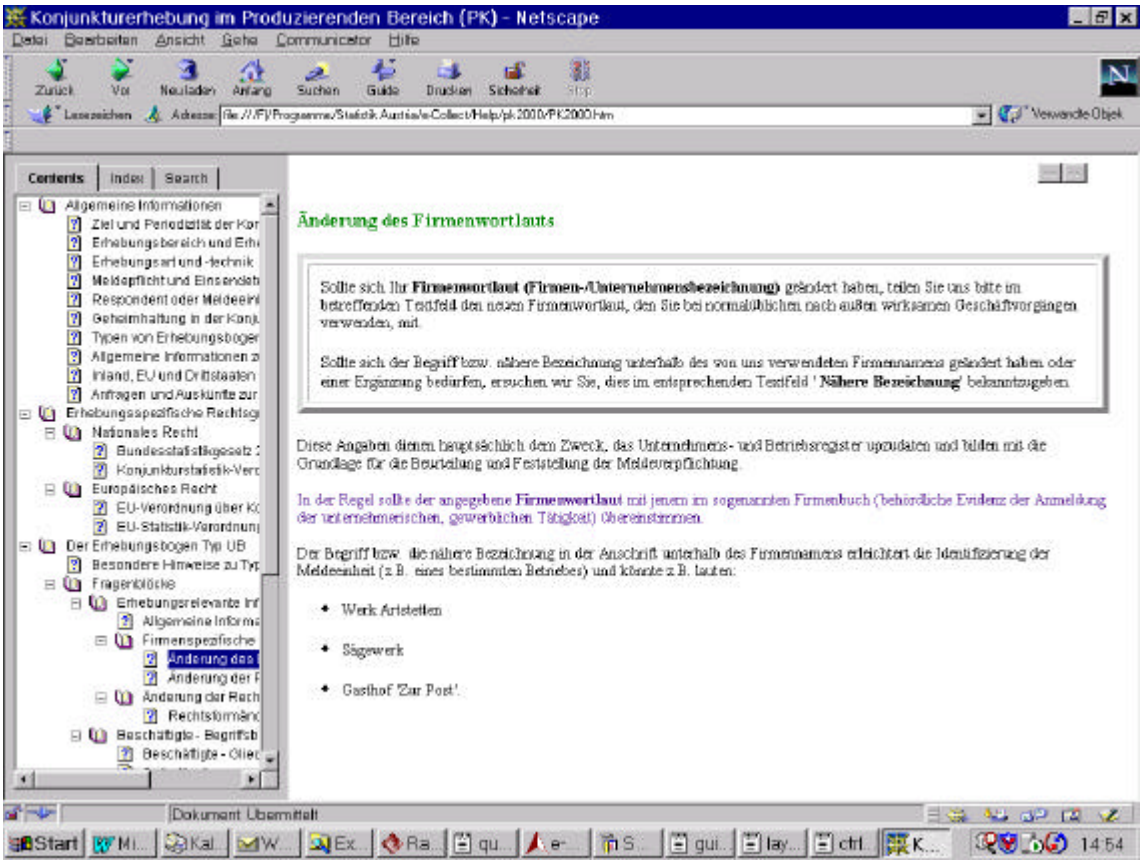

Fig. 10

52. Metadata layer 4 comprises general information about the survey, e.g. its objectives, its periodicity, its sampling methods, and so on. For the paper variant of short-term statistics, this information is provided in printed form. In e-Collect it is part of the HTML help system, but it is considerably more detailed. The complete text of the European Union's regulations on short-term statistics as well as those of the responsible Austrian ministry and the Federal Statistics Act 2000 will be included for those respondents who are interested in the legal basis of the survey. It will be no problem to supplement this material with further documents and to provide (for example) general information about Statistics Austria or results of previous surveys.

53. Classifications are a special type of metadata designed for respondents. While some classifications are just short code lists (like "Gender" with the items "Male" and "Female"), others are quite complex and hierarchically structured "information objects" available in different versions. Elements may be linked to elements of other versions and other classifications or to metadata (e.g. a list of synonyms allocated to classification items). With paper questionnaires where specific codes have to be entered (like PRODCOM codes for the classification of produced goods), dozens of pages describing the classification items sometimes have to be provided to the respondents. Obviously, such voluminous classifications do not facilitate the job of completing questionnaires.

54. During analysis for e-Collect, we had some difficulties at first in finding a solution to the problem of including classifications in the software. Short code lists are no problem, they can be defined within the questionnaire XML (as can be seen in the excerpt above). But how to cope with extensive classifications – particularly as e-Collect should be usable for different surveys, which in turn may have different requirements regarding classifications? It did not seem feasible to develop a generalised object model for classifications in the short period of time available for the SDSE project.

55. The idea we finally came up with was to out-source classifications into separate COM components which can be deployed and installed later. For communication between e-Collect and such a component, the following COM interface IEqmsClassification has been defined:

```
interface IEqmsClassification: IDispatch {
  HRESULT CreateContext([in, out] BSTR* strSurveyID,
    [in, out, optional] BSTR* strSurveyVersionID,
    [in, out, optional] BSTR* strSurveyInstanceTime,
    [in, out, optional] BSTR* strClassificationName,
    [in, out, optional] BSTR* strLanguage,
    [in, out, optional] BSTR* strFilter,
    [in, out, optional] IDispatch** oEqmsInterface,
    [out, retval] short* );
  HRESULT FreeContext([in, out] short* nContextID);
  HRESULT GetClassification([in, out] short* nContext,
    [in, out] long* hWnd,
    [in, out] VARIANT_BOOL* blnShow,
    [in, out] VARIANT_BOOL* blnShowModal,
    [in, out] VARIANT* vAdditionalOutput,
    [out, retval] BSTR* );
  HRESULT CheckCode([in, out] short* nContext,
    [in, out] BSTR* strCode,
    [out, retval] VARIANT_BOOL* );
  HRESULT CheckAdditionalOutput([in, out] short* nContext,
    [in, out] BSTR* strCode,
    [in, out] BSTR* strName,
    [in, out] VARIANT* vValue,
    [out, retval] VARIANT_BOOL* );
  HRESULT GetAdditionalOutput([in, out] short* nContext,
    [in, out] BSTR* strCode,
    [out, retval] VARIANT* );
};
```

56. This architecture provides us with a lot of flexibility. It is possible to specifically develop a component according to the requirements of a certain survey and of a certain classification. This is what was performed for short-term statistics, for which two components of PRODCOM – including extensive PRODCOM-specific metadata and links to codes of the Combined Nomenclature – and of NACE were programmed by CSC Austria. It would be possible, however, to write a more general component which could be used for several classifications.

57. At the moment all classification data and metadata are stored within the RDBMS, which is part of e-Collect, but this is no mandatory design. At some time in the future, a component will perhaps be capable of querying a (not yet existent) classification server of Statistics Austria or even of leaving the choice to the users whether the classification data should be accessed online or downloaded for local storage. If the IEqmsClassification interface is supported there will be no need to change the e-Collect source code to integrate such an advanced component into the system.

58. There is yet another benefit of the COM architecture: the classification components can be re-used within other software, for example, in data editing programs of Statistics Austria where the statisticians must be able to search for classification codes.

59. A last type of e-Collect metadata designed for respondents is not aimed at users filling in the statistical declarations, but at IT experts. It is possible to export questionnaires so that programmers can apply these files as a model when they develop extensions to the respondent's EDP systems which will automatically generate import files for e-Collect. There are options to include short explanations of the XML tags and further information (which is specified in the questionnaire XML) in those model files. Here is an example:

```
<?xml version="1.0" encoding="iso-8859-1"?>
<EXPORT>
  <Exportpack ValidDate="20001030">
    <ExportStructdata></ExportStructdata>
    <ExportXML>
      <QuestionnaireExport CSI_ID="stat.at" RSP_ID="0004711" SVY_ID="DK"
      QUT_ID="U" OBU_ID="12345678" SVV_Version="2000" SVI_Number="12"
      OBU_ValidFrom="20000101" RSP_ValidFrom="20000101"
      QUE_ValidDate="20001201" quName="U">
        <Question name="neuName" valueType="string" canBeChanged="yes"
        valueRequired="no" valuesFree="yes">
          <Answer>Max Mustermann</Answer>
          <ShortDescription>Neuer Firmenwortlaut</ShortDescription>
        </Question>
        <Question name="firmZ2" valueType="string" canBeChanged="yes"
        valueRequired="no" valuesFree="yes">
          <Answer>Hauptfiliale</Answer>
          <ShortDescription>Nähere Bezeichnung der Firma</ShortDescription>
        </Question>
        <Question name="standortStr" valueType="string" canBeChanged="yes"
        valueRequired="no" valuesFree="yes">
          <Answer>Mustergasse 19a</Answer>
          <ShortDescription>Straßenadresse des neuen
          Standorts</ShortDescription>
        </Question>
        <Question name="standortPLZ" valueType="int" maxLength="4"
        canBeChanged="yes" valueRequired="no" valuesFree="YES">
          <Answer>1111</Answer>
          <ShortDescription>Postleitzahl des neuen
          Standorts</ShortDescription>
        </Question>
        ...
      </QuestionnaireExport>
    </ExportXML>
  </Exportpack>
</EXPORT>
```

60. So much for metadata designed for respondents. With regard to metadata used by the software, two types can be distinguished as was already mentioned in chapter II:

- structural metadata (as we call them) defining the collector of statistical information, the survey and its versions, the types of observation units and their respective questionnaire types, hierarchical relationships between observation unit types, and some other objects; and
- respondent-specific metadata describing the respondent, the actual observation units for which the respondent must fill in questionnaires, the actual relationships between those units, and so on.

61. These XML metadata are deployed in encrypted and compressed form in CAB files, which can be put on CD-ROMs, sent on floppy discs or via e-mail or, later on, will be downloaded from the world wide web. For access to respondent-specific metadata files (which may include statistical data imported automatically whenever a new questionnaire is generated – the so-called initialization data), every respondent requires an ID and an encryption key. Structural metadata, on the other hand, are not really confidential so the key necessary for decoding them is hardcoded in e-Collect.

62. For modelling the SDSE, the Unified Modeling Language (UML) was used. Some of the classes identified for e-Collect are:

- CCollStatInfo: the collector of statistical information carrying out 1 – n surveys

- CSurvey: a statistical survey can consist of 1 – n survey versions
- CSurveyVersion: particular version of the survey, valid for a certain period of time (e.g. short-term statistics 2001)
- CObservationUnitType: type of observation unit, for example “enterprise“ or “establishment“
- CHierarchyRelationType: possible hierarchical relationship between observation units, specified by two observation unit types
- CQuestionnaireType: type of questionnaire, referring to exactly one observation unit type
- CHierarchyGroup: set of possible hierarchical relationships between observation units
- CSurveyInstance: actual instance of a survey, e.g. short-term survey for January 2001
- CSurveyPlan: rules for generating periodical survey instances
- CRespondent: a respondent can be responsible for one or more surveys
- CObservationUnit: observation unit.

63. From these and other classes, the XML tags for metadata files as well as the data model for the RDBMS were derived. The database for e-Collect consists of 36 tables, including those tables that are used for storing questionnaire data and metadata in XML format and tables needed for the user authorization features of the software.

ANNEX 1

Exerpt from an XML questionnaire

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<STRUCTDATA>
  <StructDataPack>
    ...
    <QuestXML CSI_ID="stat.at" SVY_ID="DK" QUT_ID="U" SVV_Version="2000">
      <QUDEF>
        <QuestionnaireDefinition CSI_ID="stat.at" SVY_ID="DK" QUT_ID="U"
        SVV_Version="2000" quName="DK 2000 U" caption="Konjunkturerhebung
        2000 – Handel und Dienstleistungen - U">
          ...
          <ItemGroupDef itemName="RForm" helpContextID="RFA"
          caption="Änderungen der Rechtsform des Unternehmens"
          shortDescription="Rechtsform des Unternehmens mit Änderungsmöglichkeit">
              <ElementaryItemDef itemName="RF" caption="" valueType="int"
            shortDescription="Code der derzeit bekannten Rechtsform des Unternehmens"
            initVisible="no" initActive="no" questionType="STAT"
            canBeChanged="no"></ElementaryItemDef>
              <ElementaryItemDef itemName="rfAlt" caption="Rechtsform derzeit
            registriert als:" valueType="string" shortDescription="Derzeit bekannte
            Rechtsform des Unternehmens" initActive="no" questionType="STAT"
            canBeChanged="no"></ElementaryItemDef>
              <ElementaryItemDef itemName="rfAnders" caption="Rechtsform nicht
            zutreffend, bitte ändern:" valueType="boolean" shortDescription="Checkbox
            zur Anzeige, dass sich die Rechtsform geändert hat"
            initValue="0"></ElementaryItemDef>
              <ElementaryItemDef itemName="rfNeu" helpContextID="RFN"
            caption="Richtige Rechtsform des Unternehmens" valueType="string"
            shortDescription="Neue Rechtsform auswählen oder angeben"
            initVisible="no" valuesFree="yes">
              <EnumValue enumCode="01">nicht protokollierte
              Einzelfirma</EnumValue>
              <EnumValue enumCode="09">protokollierte
              Einzelfirma</EnumValue>
              <EnumValue enumCode="02">Ges.n.b.R. (Gesellschaft nach
              bürgerlichem Recht)</EnumValue>
              <EnumValue enumCode="03">OHG (offene
              Handelsgesellschaft)</EnumValue>
              <EnumValue enumCode="04">KG
              (Kommanditgesellschaft)</EnumValue>
              <EnumValue enumCode="11">KEG (Kommandit
              Erwerbsgesellschaft)</EnumValue>
              <EnumValue enumCode="10">OEG (offene
              Erwerbsgesellschaft)</EnumValue>
              <EnumValue enumCode="00">Ges.m.b.H.u.Co.KG</EnumValue>
              <EnumValue enumCode="05">Ges.m.b.H. (Gesellschaft mit
              beschränkter Haftung)</EnumValue>
              <EnumValue enumCode="06">AG (Aktiengesellschaft)</EnumValue>
              <EnumValue enumCode="07">Genossenschaft, Reg.Gen.,
              Reg.Gen.m.b.H.</EnumValue>
              <EnumValue enumCode="08">Sonstige (z.B.: Verein (privater),
              öffentl. Unternehmungen,...): bitte anführen!</EnumValue
```

```
        </ElementaryItemDef>
      </ItemGroupDef>
      <ItemGroupDef itemName="Beschaeftigte" caption="1. Zahl der
      Beschäftigten insgesamt zum Ende des Berichtsmonats"
      valueType="string" shortDescription="">
            <ElementaryItemDef itemName="EBESCH" helpContextID="BESCHH"
        caption="" valueType="int" minvalue="0" maxvalue="99999"
        shortDescription="Zahl der Beschäftigten insgesamt"
        checkcode="TXVX"></ElementaryItemDef>
      </ItemGroupDef>
      <ItemGroupDef itemName="Umsatz" caption="2. Gesamtumsatz im
      Berichtsmonat" shortDescription="">
            <ElementaryItemDef itemName="EUMSATZ" helpContextID="GESU"
        caption="(ohne Umsatzsteuer)" valueType="int" minvalue="0"
        maxvalue="999999999" shortDescription="Gesamtumsatz im Berichtsmonat"
        checkcode="TXVX"></ElementaryItemDef>
            <ElementaryItemDef itemName="EUMSATZATS"
        helpContextID="GESU" caption="" valueType="int" minvalue="0"
        maxvalue="13759999" shortDescription="Gesamtumsatz im Berichtsmonat"
        checkcode="TXVX"></ElementaryItemDef>
            <ElementaryItemDef itemName="Wauswahl" helpContextID="WINFO"
        caption="Wert in:" valueType="string" valuesfree="no" initVisible="yes"
        shortDescription="Auswahl der verwendeten Währung">
              <EnumValue enumCode="1">1 EURO</EnumValue>
              <EnumValue enumCode="2">1.000 Schilling</EnumValue>
            </ElementaryItemDef>
      </ItemGroupDef>
    </QuestionnaireDefinition>
</QUDEF>
<GUIDEF>
    <GUIDefinition CSI_ID="stat.at" SVY_ID="DK" QUT_ID="U"
    SVV_Version="2000">
      <FontDef fontID="fontDefault" fontName="MS Sans Serif"
      fontBold="no" fontItalic="no" fontSize="10"></FontDef>
      <FontDef fontID="fontSmall" fontName="MS Sans Serif"
      fontBold="no" fontItalic="no" fontSize="8"></FontDef>
      <FontDef fontID="fontSmallBold" fontName="MS Sans Serif"
      fontBold="yes" fontItalic="no" fontSize="8"></FontDef>
      ...
      <GroupCtlDef itemID="RForm">
        <CtlDefFrame ctlType="Frame" fontID="fontbold"></CtlDefFrame>
      </GroupCtlDef>
      <ItemCtlDef itemID="RForm.rfNeu">
        <FieldCtlDef>
          <CtlDefComboBox ctlType="RBGroupText" toolTipText="Rechtsform
          anklicken oder angeben" Behavior="1" ></CtlDefComboBox>
        </FieldCtlDef>
      </ItemCtlDef>
      <ItemCtlDef itemID="RForm.rfAnders">
        <FieldCtlDef>
          <CtlDefCheckBox ctlType="CheckBox" fontID="fontdefault"
          ></CtlDefCheckBox>
        </FieldCtlDef>
      </ItemCtlDef>
      <ItemCtlDef itemID="RForm.rfalt">
        <FieldCtlDef>
```

```
          <CtlDefLabel ctlType="TextBox" borderStyle="0"
          fontID="fontbold10" foreColor="colButtonText"
          backColor="colButtonFace"></CtlDefLabel>
        </FieldCtlDef>
        <CaptionCtlDef>
          <CtlDefLabel ctlType="Label" ></CtlDefLabel>
        </CaptionCtlDef>
      </ItemCtlDef>
      <GroupCtlDef itemID="Beschaeftigte" >
        <CtlDefFrame ctlType="Frame" fontID="fontbold10"></CtlDefFrame>
      </GroupCtlDef>
      <ItemCtlDef itemID="Beschaeftigte.EBESCH">
        <FieldCtlDef>
          <CtlDefTextBox ctlType="TextBox" ></CtlDefTextBox>
        </FieldCtlDef>
      </ItemCtlDef>
      <GroupCtlDef itemID="Umsatz" >
        <CtlDefFrame ctlType="Frame" fontID="fontbold10"></CtlDefFrame>
      </GroupCtlDef>
      <ItemCtlDef itemID="Umsatz.EUMSATZ">
        <FieldCtlDef>
          <CtlDefTextBox ctlType="TextBox" ></CtlDefTextBox>
        </FieldCtlDef>
        <CaptionCtlDef>
          <CtlDefLabel ctlType="Label" ></CtlDefLabel>
        </CaptionCtlDef>
      </ItemCtlDef>
      <ItemCtlDef itemID="Umsatz.EUMSATZATS">
        <FieldCtlDef>
          <CtlDefTextBox ctlType="TextBox" ></CtlDefTextBox>
        </FieldCtlDef>
      </ItemCtlDef>
      <ItemCtlDef itemID="Umsatz.Wauswahl">
        <FieldCtlDef>
          <CtlDefComboBox ctlType="RBGroup" ></CtlDefComboBox>
        </FieldCtlDef>
      </ItemCtlDef>
    </GUIDefiniition>
</GUIDEF>
<LAYOUTDEF>
    <LayoutDefinition CSI_ID="stat.at" SVY_ID="DK" QUT_ID="U"
    SVV_Version="2000">
      ...
      <TabDef tabID="tabInfo" caption="Info/Änderung"></TabDef>
      <TabDef tabID="Tabdaten" caption="Daten"></TabDef>
      ...
      <CtlLayoutDef itemID="RForm" left="120" top="9525" height="6255"
      width="9495" tabID="tabInfo"></CtlLayoutDef>
      <CtlLayoutDef itemID="RForm.rfNeu" left="120" top="1540"
      height="4575" width="9135"></CtlLayoutDef>
      <CtlLayoutDef itemID="RForm.rfAnders" left="120" top="940"
      height="495" width="5415"></CtlLayoutDef>
      <CtlLayoutDef itemID="RForm.rfalt-caption" left="120" top="580"
      height="375" width="3500"></CtlLayoutDef>
      <CtlLayoutDef itemID="RForm.rfalt" left="3800" top="580"
      height="375" width="5455"></CtlLayoutDef>
      <CtlLayoutDef itemID="Beschaeftigte" left="120" top="240"
      height="1455" width="9975" tabID="tabDaten"></CtlLayoutDef>
      <CtlLayoutDef itemID="Beschaeftigte.EBESCH" left="480" top="480"
      height="375" width="1215" alignment="1"></CtlLayoutDef>
```

```
      <CtlLayoutDef itemID="Umsatz" left="120" top="2400" height="2295"
      width="9975" tabID="tabDaten"></CtlLayoutDef>
      <CtlLayoutDef itemID="Umsatz.EUMSATZ" left="480" top="600"
      height="375" width="1215" alignment="1"></CtlLayoutDef>
      <CtlLayoutDef itemID="Umsatz.EUMSATZATS" left="480" top="600"
      height="375" width="1215" alignment="1"></CtlLayoutDef>
      <CtlLayoutDef itemID="Umsatz.EUMSATZ-Caption" left="480"
      top="240" height="375" width="3855"></CtlLayoutDef>
      <CtlLayoutDef itemID="Umsatz.Wauswahl" left="3000" top="600"
      height="1095" width="4000"></CtlLayoutDef>
    </LayoutDefinition>
  </LAYOUTDEF>
  <CTRLDEF>
    <QuestionnaireCtrlDef CSI_ID="stat.at" SVY_ID="DK" QUT_ID="U"
    SVV_Version="2000">
      <VarDef VarName="saveNeuName" VarType="string"></VarDef>
      <VarDef VarName="saveFirmZ2" VarType="string"></VarDef>
      <VarDef VarName="saveStandortstr" VarType="string"></VarDef>
      <VarDef VarName="saveStandortplz" VarType="string"></VarDef>
      <VarDef VarName="saveStandortGem" VarType="string"></VarDef>
      <VarDef VarName="saveAenddat" VarType="string"></VarDef>
      <EventAction eventType = "afterInit">
        <CodePiece>init_AdrInfo</CodePiece>
        <CodePiece>layout_Info</CodePiece>
        <FieldPattern>Form</FieldPattern>
      </EventAction>
      ...
      <SubDef subName="init_AdrInfo">
        &apos; initialisiere Adressinfomation
        setVal "adrInfo.neuName",val("kopf.FAName.F1")
        setVal "adrInfo.firmz2",val("kopf.FAName.F2")
        setVal "adrInfo.standortstr",val("kopf.FAName.adresse")
        setVal "adrInfo.standortplz",val("kopf.FAName.plz")
        setVal "adrInfo.standortgem",val("kopf.FAName.postamt")
        setVal "adrInfo.aenddat",currentDate
        layout_Info
      </SubDef>
      ...
      <SubDef subName="layout_Info">
        &apos; layout für info-Tag
        if propBool("adrInfo.neuName","visible") then
          setProp "adrInfo","height",4600
          setProp "RForm","top",9525
        else
          setProp "adrInfo","height",1280
          setProp "RForm","top",6205
        end if
        if valBool("RForm.rfanders") then
          setProp "RForm","height",6255´
        else
          setProp "RForm","height",1500
        end if
        setProp "Form","AdjustTabFrames",0
      </SubDef>
      ...
    </QuestionnaireCtrlDef>
  </CTRLDEF>
  <PLAUSDEF>
    <QuestionnairePlausDef CSI_ID="stat.at" SVY_ID="DK" QUT_ID="U"
    SVV_Version="2000" shortDescription="">
```

```
            <PlausAction eventType="PLAUS">
              <CodePiece>ALLCHK</CodePiece>
              <FieldPattern>FORM</FieldPattern>
            </PlausAction>
            <PlausCheck checkName="EBESCH" priority="R" errorText="Im Merkmal
            &quot;Zahl der Beschäftigten insgesamt zum Ende des
            Berichtsmonats&quot; wurden keine Daten eingetragen. Wenn die
            Meldeeinheit zum Ende des Berichtsmonats keine Beschäftigten
            hatte, dann ist im Eingabefeld &quot;0&quot;
            einzugeben."></PlausCheck>
            <PlausCheck checkName="EUMSATZ" priority="R" errorText="Im
            Merkmal &quot;Gesamtumsatz im Berichtsmonat&quot; wurden keine
            Daten eingegeben. Wenn im Berichtsmonat kein Umsatz angefallen
            ist, dann ist im Eingabefeld &quot;0&quot;
            einzugeben."></PlausCheck>
            <SubDef subName="ALLCHK">
              setPlausStatus "Beschaeftigte.EBESCH","EBESCH",
              not(val("Beschaeftigte.EBESCH")=""), "Beschaeftigte.EBESCH"
              setPlausStatus "Umsatz.EUMSATZ","EUMSATZ",
              not(val("Umsatz.EUMSATZ")=""),
              "Umsatz.EUMSATZ,Umsatz.EUMSATZATS"
            </SubDef>
          </QuestionnairePlausDef>
        </PLAUSDEF>
      </QuestXML>
    </StructDataPack>
</STRUCTDATA>
```