

NATIONS UNIES

COMMISSION ECONOMIQUE  
POUR L'EUROPE

ОБЪЕДИНЕННЫЕ НАЦИИ

ЭКОНОМИЧЕСКАЯ КОМИССИЯ  
ДЛЯ ЕВРОПЫ

UNITED NATIONS

ECONOMIC COMMISSION  
FOR EUROPE

---

SEMINAIRE

СЕМИНАР

SEMINAR

---

STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN  
STATISTICIANS



CES/SEM.35/SV/4  
15 July 1996

Original: ENGLISH

Seminar on Official Statistics - Past and Future  
(Lisbon, Portugal, 25-27 September 1996)

**SESSION 5: OUR LEGACY TO FUTURE GENERATIONS**

**META, THE TOOL  
HOW TO DEAL WITH THE DATA DELUGE**

Report submitted by Statistics Netherlands 1/

INTRODUCTION

1. Demand and supply of statistical information are growing fast. The development of information technology gives the opportunity to store huge amounts of data. Technically it is possible to access this information effectively. Now the problem is how we will manage to lead our clients through this information. The answer for the user walking through our warehouse stuffed with data of course would be to supply him with a map, with information on our products, with information on information. "Information on information" is a good working definition of meta-information or meta-data or just meta.

2. A further and more difficult problem is how we will reach and maintain coherence within our statistical information. For indeed, it is coherence

---

1/ Prepared by Lydia A. van der Hulst, Leon F.M. Vermeulen,  
Winfried F.H. Ypma, Statistics Netherlands .

GE.96-

between statistical descriptions of different phenomena that offer national statistical offices the possibility of an advantage above other suppliers of statistical information. That coherence of data is reflected in its meta-data. It will be the same or compatible.

3. It is clear that this paper will deal with meta-data. However, the conviction is growing that it is inadequate to regard meta-information as no more than an appendix to our regular output. It will not do to set up the meta-information no sooner than after the regular statistical production process has been completed. We should aim at explicitly laying down the meta-information of each stage of the production process. That will not only guarantee the best result for the description of the end result - meta-information of statistical output - but will also allow meta to fulfil other functions.

4. On the other hand, this call for "metafication" should not be interpreted as the intention to construct just one and only complete set of meta-data of the full statistical process. Here too the approach should be pragmatic and not an attempt to build one grand cathedral. Meta should be realised as a part of the tools statisticians use in their production process. We should start to set up meta simultaneously in different stages of the process and worry later about links. Furthermore, we should set priorities within the meta-data. Not everything at once.

5.. This paper will therefore attempt to outline a somewhat broader view on meta as a tool for more stages in the process than only the output stage. After that we will give three examples of the implementation of meta within Statistics Netherlands. Of course there is the meta on the output side. There we see the tool STATLINE that eventually should be the source of our supply of information to our clients. Here meta plays its most traditional role. The second example will be meta in our statistical tool BLAISE. It will be described how meta can be maintained from questionnaire to published table. The last example will be one regarding meta on the input-side of the process. It concerns an EDI project among enterprises.

#### WHAT IS META-INFORMATION?

6. This paper will not go into the theory behind meta-information. Others have done so before. It is however useful to say more about how we see meta information. It makes it easier later to discuss the possibilities of meta. Meta is information on information, statistical information in our case. Sundgren [1992] gives ample discussion of the issues at stake.

7. It may be useful to make some distinctions within meta. When describing statistical data we can discern two aspects.

#### Conceptual meta

8. Conceptual meta describes the contents of the data. It tells us what are we talking about when we use particular statistical data. It deals with definitions of concepts and with the classifications used to give a further breakdown of the phenomena at hand.

9. This information places the outcome on the right place (row, column) in the right table. It is also this information that should lead the user to the right kind of statistical information.

10. A special part of conceptual meta are the computational features of the data. Computational meta deals first with the units of the data (1\$ or 1000 Mbytes etc.). Second it describes whether we are dealing with a straightforward number or with an average, a price, an index, what kind of index etc. Computational meta therefore indicates how to manipulate the statistical information itself. It is essential as soon as the user actually wants to use the data e.g. for aggregation or a regression analysis.

#### Process meta

11. Process meta describes how the statistical information is produced. It describes for instance the sample methodology. Process meta is in the first place relevant for the more professional user of statistical information among which other statisticians of course. It gives insight in the quality and comparability of the data.

12. Process meta is more complex than conceptual meta. It is therefore probably more difficult to define a universal format. Furthermore, defining the relations (of comparability e.g.) between different processes will prove much more difficult than those between different concepts and classifications.

13. Especially where conceptual meta is involved one should be aware of the fact that the semantics used there always have a contextual meaning. If one were to create, to start with, separate sets of meta-information the result would be meta with only local relevance. Only definitions applicable throughout the whole of the statistical office could be called global meta (within the SO!). The solution could be to strive for unification of all meta. First however it should be considered what the cost benefit ratio would be. Second, sometimes a statistical office is forced to work with data from other sources and therefore with meta-data over which it has no say. Those data originating from outside sources will always belong to a different world. That implies that along with these data comes a separate set of meta data. Third, we find a parallel situation on the output side. Different clients often want to receive data conform different definitions i.e. with its own, local, meta.

14. The most pragmatic solution is to allow different local meta-data to exist but to create links, translations between them. Those links will not regard all but only the main data-elements.

#### **FUNCTIONS OF META INFORMATION**

15. In this section we discuss the different uses of meta information. The first function of course is the description of data. This description can be

extended to a map through or a catalogue of a collection of data. This can be done by creating thematic links between the conceptual meta we mentioned in the previous section. Besides thematic links one of course can create a thesaurus and lists of synonyms to facilitate further searching through the data.

16. Sundgren [1995] gives an example of a format for the full description of a statistical survey. It has all the elements of meta, conceptual as well as process, as mentioned above. It is intended as a mere description. There is a remarkable correspondence to the scheme presented by Koeijers & Willeboordse. The latter however are presenting a methodology, i.e. meta ex ante. So instead of descriptive meta can also be used prescriptive. Meta can and should be used to control the statistical process.

17. Here one probably first would think about process meta. But especially in the area of conceptual meta progress can be made. We mentioned earlier that our users not only ask for easy accessible but also for coherent information. That coherence can best be controlled by conceptual meta. It is here that a link can be made with statistical co-ordination or as others call it the integration of statistical concepts. Data are co-ordinated when they can be used within the same theory, the same system. This is reflected in the relation of the meta-information of the different data-items. Co-ordination regards conceptual as well as process meta. The meta should be compatible within the theoretical framework where the data-items are used. The question whether data are co-ordinated can only be answered by looking at their meta.

18. Here then, meta can, in its descriptive state, be used to find out about the degree of statistical co-ordination. Since co-ordination is mostly regarded as a desirable feature of our data we can also try to use meta for reaching co-ordination. It is co-ordination that leads to coherence.

19. This of course leaves one question. In principle there are different theoretical frameworks possible within which we could strive for statistical co-ordination. Which should we choose?

20. By all this we do not mean to say that up until now statistics were made without meta. The problem is that the meta used often was not made explicit. Only at the output stage something was done about meta. That is not enough.

#### **META AND THE STATISTICAL PRODUCTION PROCESS**

21. Within the "new" statistical production process we discern an input, a throughput and an output-phase. In the previous sections we spoke about several aspects of meta and of its functions. It is possible to be more specific about meta in the different stages of the process.

##### Output phase

22. We start at the end of the process. The reason is clear. It is here in the first place that we are looking for meta. The main user of meta here is the client of the SO. Here meta's most important function is to lead the user to the data he needs. What we need most is conceptual meta. It should be extended with user friendly search facilities etc.

23. In the second place the end-user looks for information regarding the quality of the data. Following what has been said above one thinks of process meta. Besides that however there is very relevant information on information for the user that emerges as the result of the process. Examples are confidence ratios and non-response data. (This most certainly is not meta that could be described ex ante. Ex ante one can only discuss how to deal with non response.)

#### Input phase

24. Here the main user is the respondent. The statistician has to make clear what he wants to know: what is meant by "net turnover" in the questionnaire, how to deal with VAT and should the answer refer to guilders or to 1000 guilders? What is needed in the first place therefore is conceptual meta. The problem of course is linking the local meta-system of the respondent to that of the SO. Of course we have to address the respondent as much as possible in his own language using familiar concepts.

25. Process meta here is relevant for the statistician. At the input phase we are dealing with different surveys that can be described by for instance the survey documentation template of Sundgren we already referred to.

#### Throughput phase

26. In the throughput phase statistical data is translated, transformed, combined and integrated to reach the different products of the SO. Sometimes the process is quite straightforward. There are still many surveys that pass from input directly to output, i.e. from questionnaire to published table. Here linking conceptual input meta to output meta should be fairly simple.

27. Also describing the process meta seems not to difficult. Complex processes like the one that results in the national accounts give us more problems. Linking conceptual meta still should be possible even in a formalised way. Describing the process has of course been done occasionally. The existing descriptions are however still a long way from e documentation template and cannot be used in a formalised way e.g. to control the process.

28. As said before, we advocate a pragmatic approach. This leads to the conclusion that the collection of meta in the input and the output-phase will be given the priority. Furthermore conceptual meta will be given the priority above the process meta.

#### **THE IMPLEMENTATION OF META**

29. It is not enough to reach the conclusion that meta information can control the statistical production process and lead it into the right direction. The more difficult question is how to give meta this role it should have.

30. The first possibility of course is to decree that meta is an essential (by-)product of the statistical office because otherwise the customers are left unsatisfied. There are two problems. Most statisticians tend to regard their work as done once the figures are out. They are willing to write a description of what they do and how they do it once in a while. That is a lot of work. Moreover, when something changes, the definition of a variable e.g., there is the danger that it will remain unnoticed because the meta information is not adapted. The other danger is that this practice will give no incentive for coherence. Coherence demands as we have seen identical or linked meta-information.

31. The next possibility is that centrally rules are made up describing what is to be made and how it should be done. In a decentralised organisation as Statistics Netherlands however rules are not always effective. Besides they have the disadvantage that the obedience to the rules should be checked.

32. In our view meta can only become effective when it is embedded within the tools that are used in the statistical production process. This in the first place demands that statistician will use those tools. So the tools should be attractive enough. Second, the tools should be such that they indeed can, through their meta-information, control the production process. Here the developments in the organisation of the statistical production process help. The old way was, slightly exaggerated, that for each statistic to be published a separate survey was set up addressing all the relevant sources. "Of course" each survey had its own meta. The new way will be that each source will only be addressed once and that the results will be stored in an input database. Therefore, at the input-side the walls between the separate surveys disappear. Maybe local meta-information will be stored for a specific source. From there at least a link towards the meta of the further process is essential.

33. This changes the relationship with our respondents. The old way was that they are confronted with many different surveys that only in the ideal situation used the same meta. In the new way they are confronted with only one combined questionnaire with only one set of meta-information. What will disappear is the annoying practise that different surveys asked for related but still different concepts.

34. This all does not mean that the problems of coherence or statistical co-ordination within the organisation are solved. We have in a way centralised the decision on meta, on concepts. The next question within the organisation will of course be: "Who will be the keeper of the Meta?" Within Statistics Netherlands this question has as yet not been answered.

## ILLUSTRATIONS

### Meta-data in Statline

35. Statline is Statistics Netherlands' tool for the dissemination of data. In Statline we store our data. Statline can become our archive. Statline can store our legacy to the future.

36. Statline is an output-database with all the necessary software to search for and retrieve information. It mainly contains data by means of figures but also press releases etc. For the sake of searching and retrieving it also contains a lot of descriptive meta-data.

### Functions of Statline

37. Direct access for end users to the data of SN either on line or by means of CD-ROM.

- Instrument for the information desk.
- Source for data to be distributed by publications, fax or E-mail

38. Eventually all output of Statistics Netherlands (SN) should go through the channel of Statline. Immediately we see here the possibilities for control of the structure and the contents of the information through the meta information of Statline. At the same time however we have to admit that this ideal state hasn't been achieved yet.

### Present situation

39. Data is stored in Statline in so called cubicles, multi-dimensional tables. Each item is accompanied by descriptive meta through the description of the cubicle and all the rows and columns it belongs to.

40. There is the possibility of searching on keywords through all those descriptions. Furthermore Statline has a table of contents that supplies the user with a structured overview of the available data. Of course that is just one of the many way data can be structured.

41. Data can be combined into one cubicle depending on the axis of the cubicle. This demands a certain compatibility of their meta. One cannot combine data into one "regional" cubicle that use completely different regional classifications. The less statistical co-ordination we have, the more different cubicles in Statline are necessary.

42. The meta within Statline aims in the first place at making data accessible. The user should be guided to his information in the shortest possible way. The approach here too is pragmatic. The first aim now is to enter as much as possible of SN's output in Statline along with its meta-information. Furthermore much effort is needed to co-ordinate not the data

but the meta-data. Meta too has to be co-ordinated regarding its format and the use of terminology etc.

#### Bringing meta in line: statistical co-ordination

43. Although this policy of course brings us closer towards our goal it cannot be enough. What is being done is the combination of all the different publications of SN into one system. The problem is that many of these publications have their own language. Moreover, the published statistics often use their own methods and have their own concepts. In other words we see that within Statline many local meta-systems are brought together in one publication system. Within Statline this could lead to as many cubicles as there are statistics. Often this will be enough. Individual statistics may address particular users and address them in their own language. The particular language, the own local meta-data are then even a sign of user-friendliness. However, a user searching through the complete data-set is at risk of getting lost in all the different definitions. Besides the comparative advantage of a national statistical office was the coherence of the data-set it produces.

44. Statline gives us through its meta-data an opportunity to find the major flaws in the coherence of our data.

45. The next step therefore will be to bring the meta-data in line. This in fact is the statistical co-ordination of our data. That will prove to be a tremendous job. This does not imply that SN will no longer contain data aimed at particular users. There should however be a possibility to communicate on all the information in a standardised way.

#### A tool that rules?

46. Creating one language to communicate about our data may imply the unification of several meta-data sets but also defining the relations between different local meta-data sets that are allowed to serve their particular users.

47. Even having reached this goal at one time there is always the risk of diverging statistics not sticking to the rules. This risk can be met by changing the process. Presently statistical departments deliver data and meta data to Statline. By changing the process in such way that meta data precedes the data and that only data conforming to the meta-data are accepted control of the end result is possible.

48. The ultimate goal for Statline could be to use a separate meta-database. This meta-database reflects in fact the working program of SN. It contains the product-specifications of our output. Results of the statistical production process can only reach its final destination, the client, through Statline. Data will only be allowed access into Statline if they comply with the meta-data of Statline. Working like that the rule of Statline would be

very strict. We are still a very long way from that situation but it illustrates the possibilities.

#### Meta-data in Blaise

49. The statistical process consists of many phases, each of which requires its own approach to data. The data are first collected, usually by using a questionnaire. Then they are edited, merged with other data, weighted, aggregated and finally published. These different activities often include complex tasks and involve various software packages. Blaise is one such program. It is a control centre for computer assisted survey processing. The system enables the user to manage various activities in survey processing in an easy and user-friendly way.

50. In order to use different software packages, it is necessary to transfer the data between them. It is not enough to exchange data: the meta-data must also be made available for the different pieces of software. The problem is that every package has its own data description language. Transferring data from one software package to another implies making a new data description. Introducing an extra step in the process causes a growing risk of making errors. It would be much easier to make several programs work together if they all used the same meta-data language.

51. It will take a long time before this dream becomes reality. Getting every software producer to adhere to the same universal data description language will be a long, and probably frustrating process. And the all-in-one data processing software will probably never become reality. At Statistics Netherlands we have chosen for a third approach: the integrated control centre.

#### The Blaise control centre

52. Our integrated control system consists of various modules and programs all using the same meta-data system. It is possible to add any third-party software, such as analysis packages (SPSS, SAS, etc.) or database management systems and to communicate with them through translated meta-data. This means that once one has integrated a third-party program in the Blaise Control Centre, the system knows how to convert meta-data for it. Blaise is able to use the language of any piece of software to communicate with it. Once a data description in the Blaise language is available, it is a smooth process to make this description available to other packages. Maintenance activities need only be performed in Blaise. The system will take care of supplying the other packages with the updated meta-data.

53. The integrated control centre has software modules for various data processing tasks, covering the survey process from data collection and manipulation to the production of research and publication tables. One essential module is the program for computer-assisted data collection, data entry, and data editing. It is required for the creation of a data set. Additional modules may be included, such as programs for tabulation,

imputation, weighting, and analysis. Blaise is an open system, allowing its users to include or exclude software modules. It can be tuned for use in a specific environment, so that it can be used by all survey processing departments of the statistical office. This promotes standardisation and integration.

#### The Blaise language

54. The basis of this survey system is the Blaise language. In this language one can describe a large part of the meta-data necessary for the survey process. The Blaise language includes:

**Definition of survey variables.** Each variable must have an identifying name, a domain of valid values (for qualitative variables, this is a code list), a question text required to obtain values (in as many different languages as necessary), and other texts to document the variable. The documentation texts can be labels for use in tables, documentation texts for interviewer instruction, etc.

**Data model.** The data model describes the relationships between variables in terms of groups, hierarchies, and replications. For example, groups can be nested in other groups, and groups may be replicated a number of times.

**Routing instructions.** Routing instructions define the order in which, and the conditions under which the questions (variables) are asked. Such instructions see to it that only relevant questions are asked, and that irrelevant questions are skipped.

**Relationships.** Wherever relationships impose restrictions on the values of variables, these restrictions must be specified in order to be able to carry out consistency checks on the collected data. Usually, inconsistencies are caused by errors in the data. Detection and correction of such errors lead to improved data quality. For each relation a number of texts can be defined. These texts can be used for error messages during data collection or data editing, documentation purposes, etc.

**Computations.** Not every survey variable is assigned a value by asking a question. Sometimes, new variables are derived from other variables by means of expressions. Furthermore, it must be possible to replace a missing or erroneous answer by a 'synthetic' answer. Such an answer may be the result of a computation.

**Links to other files.** For some surveys it is important to establish links to data files of other, or previous surveys. This makes it possible to import information from other surveys. This facility can also be used to compare data from panel surveys. Relations between variables from different surveys can be described.

### Bascula

55. Another module of the Blaise III Control Centre is the program Bascula for adjustment weighting. After data collection, data entry, and data editing, the sample data file is usually not ready yet for computing statistics. In order to correct for a possible bias due to non-response, some form of adjustment weighting usually has to be carried out. Bascula offers several weighting techniques. In the first place, traditional post-stratification can be carried out. And if the number of empty strata is small, it can collapse strata. If there are many empty strata, or not enough population information, Bascula can carry out linear weighting, or apply iterative proportional fitting (also called multiplicative weighting, or ranking ratio estimation). This module also uses the Blaise data description.

### Abacus

56. After data collection and weighting, tables can be made with Abacus. Abacus starts by activating a table definition window in which one can define the fields to use and their place in the table. The fields can be selected from the data model with the data selector. All the meta-data defined in the data model is available to Abacus, because it uses the same meta-data file as the rest of the system. Whatever is not of interest for defining the table is ignored, such as the order in which the fields are displayed on the screen during data collection. After selecting the fields of interest, one can place them in the rows, columns and layers of the table being defined. The block and field labels defined in the meta-data can be used for row, column, layer and table headers.

### Manipula

57. In Blaise there is also a module for manipulating data, called Manipula. It is our Swiss army knife for data handling. Manipula can be used to make selections from data-files, create new variables, change the value of certain data fields, convert data from Blaise format to ASCII and back, or combine some of these possibilities. The description of the data is, of course, the same as in the other parts of the system.

### Cameleon

58. Blaise is not meant to cover all the tasks in a large survey. Suppose, after collecting data with the Data entry program, adjusting it with Bascula and massaging the data set with Manipula, one wants to use a statistical package such as SPSS or SAS for further analysis of the data. The first step for using the other software package is to make a description of the data in the language of that package. The second step is to feed it with data. After these two steps the data is ready to be analysed. Making a data description for SAS or SPSS only involves translating the data description already available in Blaise to a data description for the other package. In Blaise there is a module to convert Blaise meta-data to the meta-data of other

packages. This module is named Cameleon. With Cameleon can manipulate the meta-data and convert it to any required format. Cameleon is the module that allows Blaise to talk the languages of other software packages.

59. Converting Blaise data to ASCII data is very simple because the description of the data is already available. The Manipula set-up to convert Blaise data to ASCII Date is just 3 lines of code. It can be used for every conversion from Blaise data to ASCII. The specifications of both the ASCII file and the Blaise file refer to the same meta--data file. The physical storage may be different, but the conceptual model is the same.

60. Converting meta-data to other packages is also very easy. Cameleon only needs a Blaise meta-data file and a set of translation instructions. The translation instructions need to be defined only once for every software package, after which they can be applied to any data model defined in Blaise. To launch the translation process it is sufficient to select the Blaise meta-data file and the translation file.

61. A great benefit of this approach is that every Blaise user can write his own Cameleon instruction file for his favourite software package. But Cameleon is not just a set-up generator: it is also a meta-data manipulator which can generate tailored set-ups for specific goals. All the Blaise meta-data are available to Cameleon and can be translated to a textual description in any language. One can also make file descriptions in Pascal or C, and even reproduce the original Blaise source.

62. Here are a few examples of output that can be generated from a Blaise data model:

- Paper questionnaires
- documentation
- instruction for interviewers
- archives
- overviews
- new data models or sub-data models
- field descriptions in Pascal or C for access to ASCII files from tailor-made programs
- Internet questionnaires for use on the World Wide Web.

63. Communication of data and meta-data from program to program is one of the strong points of Blaise. Cameleon's meta-data manipulation capacities offer enhanced productivity not only for statistical data, but in any context in which the same data have to be processed by programs using different data description formats.

64. This, however, is not a final solution. It should not be the responsibility of one package to help all others to communicate. It is our hope that eventually data description will be standardised, all software will speak the same meta-data language, translations for going from one package to other would not be necessary. For all other meta manipulation such as

generating documentation, making overviews, etc. a meta manipulator will still be needed.

#### A tool that rules?

65. Blaise "rules" in so far that it manipulates data and keeps track of the corresponding meta-data from input to output. Blaise itself does not prescribe the use of one particular kind of meta. In practice it can easily help to do so. Within Statistics Netherlands a standard set of questions regarding the household has been made, the so called household box. These questions have already been translated into the right source code for a Blaise questionnaire. The individual statistician can easily import those questions into his own questionnaire. If the tool is right and if the meta is convincing then not many explicit rules are necessary to reach the necessary co-ordination.

66. Remarkable about Blaise is the fact that it handles input, throughput and output.

#### Meta-data in EDI Pilot-2

67. EDI Pilot-2 is one of the many initiatives of SN to establish an electronic link between SN and its sources of information. It directs itself towards the financial records of individual enterprises (possibly kept by commercial bookkeeping offices). For the sake of this paper we will not deal with the technique of this project. We will merely state that software is provided by SN. After installing the software and a one-off operation of defining the links between the enterprises' books and the EDI questionnaire, answering the same questionnaire for later periods should be a matter of pushing buttons unless the questionnaire and or the enterprises' books conceptually change.

68. Let us concentrate on the meta side of the project. There the goal of the project is to unite all the questions aimed at the source in question. This meant the combination of quite a few questionnaires. The result is called the combi-questionnaire. It showed however that those questionnaires also held questions not suited for this EDI project either because they were directed towards another part of the bookkeeping system or because they couldn't be found in any records and had to be answered by the respondent in person. It was also possible that the financial records were split up in several parts with the consequence that the combi-questionnaire had to be split up as well.

69. The first thing was to make sure that the right questions were sent to the right respondent which meant a reshuffling of the questions. Moreover we wanted to address respondents in their own language. An industrial firm keeps its books in a different way from a retailer. Even with the legislation on the financial reports existing in the Netherlands, the records of enterprises are far from standardised. It is clear that they all have their own (local)

set of meta-data different from the statistical meta-data of SN. So even if we want to ask for the same concept we may have to phrase the question differently for a different respondent.

70. Last but not least we aimed at as much as possible coherence between the variations of the combi-questionnaires sent to the different types of respondents

#### A central database of concepts

71. Centrally we set up a database of concepts, phenomena, features or variables that could possibly be asked from respondents. Along with the list of concepts goes a list of extensive definitions with (eventually) all the relevant includes and excludes. Within the definitions it is possible to refer to other concepts. For the user this can lead to hypertext links. It is also possible to lay down different computational relationships between concepts. One can e.g. find how value added at factor costs is to be computed from the other concepts.

72. The database contains default questions and default explanations for each variable.

73. Besides concepts the system contains classifications for those variables that require further specification e.g. foreign trade broken down by product.

#### A system of derived questionnaires

74. Making a questionnaire is a matter of point and shoot from the central database of concepts. However one can only ask for concepts present in that central database. On the other hand one can rephrase the default question and explanation to add enough "colour local" for ones respondent. There is of course the danger that rephrasing leads to redefining.

#### Linking questionnaires to respondents

75. The "respondent" is the institution that actually answers the questions. His data regard the "described unit". Sometimes this described unit does not coincide with one of our "statistical units" and a further translation is necessary. For the sake of this form of EDI a register is necessary that keeps track of these three kinds of units. Furthermore this register makes clear which unit is to receive which questionnaires.

#### Further remarks

76. One should be aware of the fact that the whole project links two local sets of meta-data. In this case we expect the respondent to make the translation from his data to those of the combi-questionnaire. This may sometimes involve a adaptation of the system of the client. It should be noted however that the contents of the questionnaire already before the

introduction of EDI developed in the direction of the main bookkeeping practices and the law on this point. This indicates that a further translation may be necessary towards other systems like those of a user further down the line, e.g. National Accounts.

77. Another point to be made is that within this system it is now fairly easy to establish the relations to other (local) meta-data-sets. We have already made an attempt to link our 700 concepts to the concepts of the Structural Business Statistics of EUROSTAT. We will also try to link our variables to those of the fiscal "questionnaires" to be filled in by our respondents.

#### A tool that rules?

78. As said before, the combi questionnaire contains the questions of several previously separate surveys. Previously the surveying departments had complete freedom in making their own questionnaire. Setting up the combi questionnaire we tried to comply with their wishes and their need for information. Adaptations were unavoidable but so far all participants could be convinced. (The exception might be the problem of those - sometimes few - questions that were not fit for this EDI project. Not everywhere the solution is clear)

79. But once lured by the EDI-sirens with its many advantages, one can only change the questionnaire conceptually if one has convinced the administrator of the database of concepts. The organisational consequences of EDI for SN are not clear yet, but it is clear that here not only a proper description of the collected data is available but also the possibility to get a grip on the contents of the data. Here the possible role of meta in controlling the process is evident.

#### **CONCLUSION**

80. There are strong similarities between the three examples given. Everywhere the first concern is: Let there be meta!. As soon as there is meta we worry about its format. We need to be able to communicate meta. Blaise has developed tools to this end. Statline is confronted with this problem as well. Blaise already is able to transfer meta through the statistical process. Neither EDI nor Statline can bother with that problem at this stage. That is the consequence of the piece-meal approach we advocate. Of course we dream of automated transfer of meta e.g. from Blaise to Statline.

81. The third phase will be to bring the meta itself in line as far as necessary. This amounts to statistical co-ordination. This problem is felt deepest by those working on Statline.

82. It also has become clear that all these tools for statisticians may help to control the process and its results. This is the last phase of the "metafication". Even if we were able to rule by the tools and reach in that

way a product that is easily accessible as well as coherent two questions remain for the SO.

What should the coherence look like?

How are we going to maintain the meta within the organisation?

---

### Reference

- BLAISE III, A survey processing system. Statistics Netherlands, 1994
- Sundgren, Bo, 1992: Statistical Meta-information Systems - pragmatics, semantics, syntactics. R&D report, Statistics Sweden.
- .., 1995: Making Statistical Data more available. Paper prepared for the ISI conference of 1995.
- Keller. W.J.; Ypma, W.F.H. 1995: Electronic Data Interchange for Statistical Data Collection, paper presented at the NTTS conference, Bonn, 1995.
- Koeijers, Elly; Willeboordse Ad, 1995: Reference manual on design and implementation of business surveys, Statistics Netherlands