**Economic and Social
Council**

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Meeting on the Management of Statistical Information Technology
(Geneva, Switzerland, 15-17 February 1999)

Topic (i): The impact of Internet on the statistical production and
dissemination process

META TOOLS IN SUPPORT OF A CORPORATE DISSEMINATION STRATEGY

Submitted by Statistics Netherlands[1]

I.   INTRODUCTION

1.   In the early nineties, Statistics Netherlands decided to begin
development of a central on-line retrievable data warehouse, StatLine. The
importance of this decision for the dissemination policy of Statistics
Netherlands cannot be overestimated, especially because this decision was
accompanied by the strategic statement that StatLine was deemed to become the
one and only source for *all* publications of Statistics Netherlands. The first
step in this process was the gradual filling of StatLine with the
publications, or rather their tables and their explanatory notes, as produced
by the wide variety of statistics. This process is now in the stage of
completion: by the end of 1998 almost all statistics will have their place in
StatLine, although still not all of them provide the full range of their data.

2.   Until recently, StatLine accepted statistics (i.e. tables) as they were
supplied, regardless of their shape, content and the quality of accompanying
documentation. Although the StatLine technology offers the possibility to

---

[1]   Prepared by Ad Willeboordse and Winfried Ypma.

store data in *multi-dimensional datacubes*, and thus to concentrate different "paper tables" in one datacube, producers of statistics often prefer to deliver their data as mere duplicates of their paper tables. Besides, a basic feature of the StatLine system is that each individual statistic has its own "guest room" in the warehouse, in which the owner of the statistic can store his datacube(s) independently from those of others. In its present state StatLine can, therefore, for the most part be considered as the set of all (paper) publications, with a few exceptions such as the regional datacube and the historical time series, in which data from a variety of surveys are combined.

3.   The contents of StatLine are presented to the users as one giant publication, both on the Internet and on a CD-Rom. This being the case, it is obvious that the content should be coherent: in a sense each datacube can be said to represent a piece of a jigsaw featuring society as a whole.

4.   When looking at StatLine as it is now, many of the pieces do not fit too well; indeed, a number of weaknesses occur in this respect. Firstly, there is simply a lack of uniformity in the presentation of data and metadata. It may be true that such cosmetic differences are statistically innocent, but nonetheless they may confuse users surfing through the warehouse. Secondly, more so than in paper publications, weaknesses in the description of the metadata become apparent, particularly with respect to definitions of concepts. Thirdly, the same terms have different meanings on different places in the StatLine. Conversely, for the same concepts different terms apply. Fourthly, and this is, statistically speaking, the most serious drawback, concepts from different statistics are incoherent and/or data are inconsistent, so that either the data are contradictory or they cannot be meaningfully related. Fifthly, the separation of data over a large number of different tables according to the borderlines between statistics is often artificial from the angle of the "area of interest" of users. Consequently, the latter have to travel around through the Warehouse to gather the data required.

5.   This paper discusses the way Statistics Netherlands deals with the se problems, or, to put it more positively, the way StatLine and its surroundings are utilized as *tools* to achieve a long-cherished goal: stimulating established coherence in concepts and data and the explicit showing of this coherence. Thus, the impact of StatLine exceeds its primary goal of an integrated presentation of data to users: it also triggers integration of concepts and data themselves. In section II we will describe the development of the tools around StatLine as the central feature in the corporate dissemination process. In this description, *meta* plays the key role. Section III then discusses the way the tools are actually used for enhancing coherence and accessibility of the package of statistics. The paper concludes with some remarks on the role of StatLine in the overall dissemination policy of Statistics Netherlands.

II.   STATISTICAL META INFORMATION TOOLS

6.   This paper focuses on the role of statistical meta information in the final stage of the statistical process. Statistical meta information describes the statistical data. The way we intend to process our data puts certain demands on the way meta data are processed. Therefore, we start with a short description of the statistical (output) process. Then we will describe the general datamodel of our metadata and this will lead us to the structure of the system.

II.1 The output phase of the statistical process

7.   The structure of the process Statistics Netherlands is now wishing to achieve looks as follows:
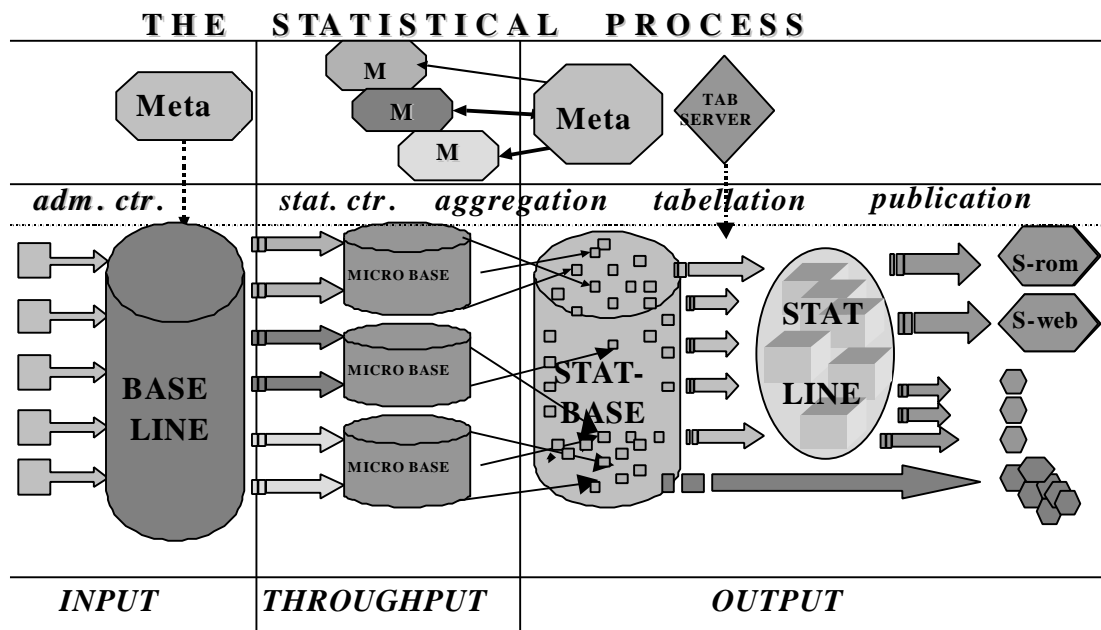
In view of the subject of this paper, we will confine ourselves to the middle and right part of the picture. The *micro databases* in the throughput phase contain the final result of the throughput process, at the level of individual data. Most important are those on individuals/households and those on enterprises/establishments. These data cannot be published for reasons of confidentiality and reliability.

8.   From the micro databases a *reference database,* StatBase, is filled by aggregation. It contains aggregated data that are reliable and safe enough to be published. Therefore, between micro and reference there is a function that aggregates in such a way that confidentiality and reliability rules are obeyed. In the reference database there is no predestined ordering of data for whatever publication. All we know is that data in the reference database *can* be used for publication; whether and in what context they are published is a

matter of dissemination policy. The reference database is technologically dedicated to this function. Updating and adding data should be easy. Fast retrieval is less important.

9.   From StatBase (and from StatBase only!) the *output datawarehouse* StatLine is filled. To this end, we will have to order and structure our information to the needs of our users. Maybe we will have to present the same information in more than one way to suit different needs. This is a specific process that demands specific skills. We used to speak of the *art of cubism*, referring to the skills needed to construct the appropriate *datacubes* needed for StatLine [see Altena and Willeboordse, 1997]. Thus, StatLine can be seen as a data warehouse containing a number of datacubes, each covering an "area of interest" and together providing a comprehensive and coherent picture of society. Furthermore, StatLine is tuned to enable easy and fast retrieval of data by our users. Updating is less easy and less fast. This is no problem: it is, after all, not a floating database.

II.2 The model of statistical meta information

10.   Around our output database we have adopted the datamodel as described by Sundgren [1992]. Much simplified, it goes as follows: Statistical data describe (sets or populations of) *objects*, i.e. instances of an object-type (e.g. enterprises). Of these objects certain *count variables* are described (e.g. turnover). The population of objects may be subdivided into subpopulations using a *classification variable* that constructs a classification (e.g. economic activity that leads to the International Standard Industrial Classification (ISIC)). This is mostly done for a certain period of *time.*

11.   Using this model, which looks simple enough, we can analyze any "typical" statistical statement. Still, there are problems. It is not easy to find the object-type behind the number of hours of sunshine and the millimeters of rainfall in November in the Netherlands. Also, the more complex and aggregated statistics are not easy to describe in terms of objects, classification variables and count variables. Consider a statement about the national saving surplus of the Netherlands in 1997. What is the relevant object-type here?

12.   Being aware of these problems, we still take the model to be applicable for all types of statistical data and throughout the whole statistical process, maybe even more obvious at the input side than at the output side. Our experience so far shows that it is not easy to implement the model in the way the average statistician looks at his or her data. Each dataset, however simple, has to be analyzed in terms of this model.

II.3 A tool for storage of data: statbase

13.   For the time being, we consider the structure suggested by the model described above as the most appropriate one to be implemented in the reference

database StatBase. Here, the information is *stored* so that – one stage further on in the process - StatLine will *order* in the form which attracts most customers.

14.   The fact that there will be (StatBase is still under construction) a database of aggregates between the micro databases and StatLine is new in itself for Statistics Netherlands. Before, statistical data had to be delivered directly, in the proper form, to StatLine.

15.   StatBase does not (explicitly) structure the data into datacubes as StatLine does. The data is grouped according to delivery or source, but this classification in no way hinders StatLine tables consisting of data from different groupings.

16.   Before data can be read into StatBase, two kinds of metadata will have to be supplied:

- The metadata describing the *contents* of the data. The format here is rather strict and, of course, dictated by the model described above. Furthermore, metadata are not delivered alongside the data. Metadata are stored in a central meta-database or *metaserver* . Data are described by reference to the metaserver. This metaserver will be described below.

- The metadata describing the technical *layout* of the dataset to be read. This is very straightforward. Having described above what information may be expected, one now indicates where and how in the dataset each item may be found.

17.   The actual way data are grouped in StatBase is irrelevant. In the beginning the functionality will be at a relatively modest level. It is obvious that there are enough possibilities to gradually increase the uses of the database.

18.   Of course it should also be possible to construct cubes for StatLine from the data stored in StatBase. Here we need the same kind of meta:

- The metadata selecting and structuring the data to be entered into the StatLine datacube.

- The metadata to be presented to the user of the StatLine cube. One needs names, labels, notes, etc.. Of course the names, labels and definitions in the metaserver will serve as defaults but these will not always suffice. The possibility should be open to provide adapted wording, etc. (at the risk, however, of being imprecise).

19.   In the actual system we find these four kinds of meta back in three separate elements: the actual database StatBase, the MetaServer, the InputServer and the TabulationServer. The actual process looks as follows:

```
⟹ `Data          ⟶ Metadata
```
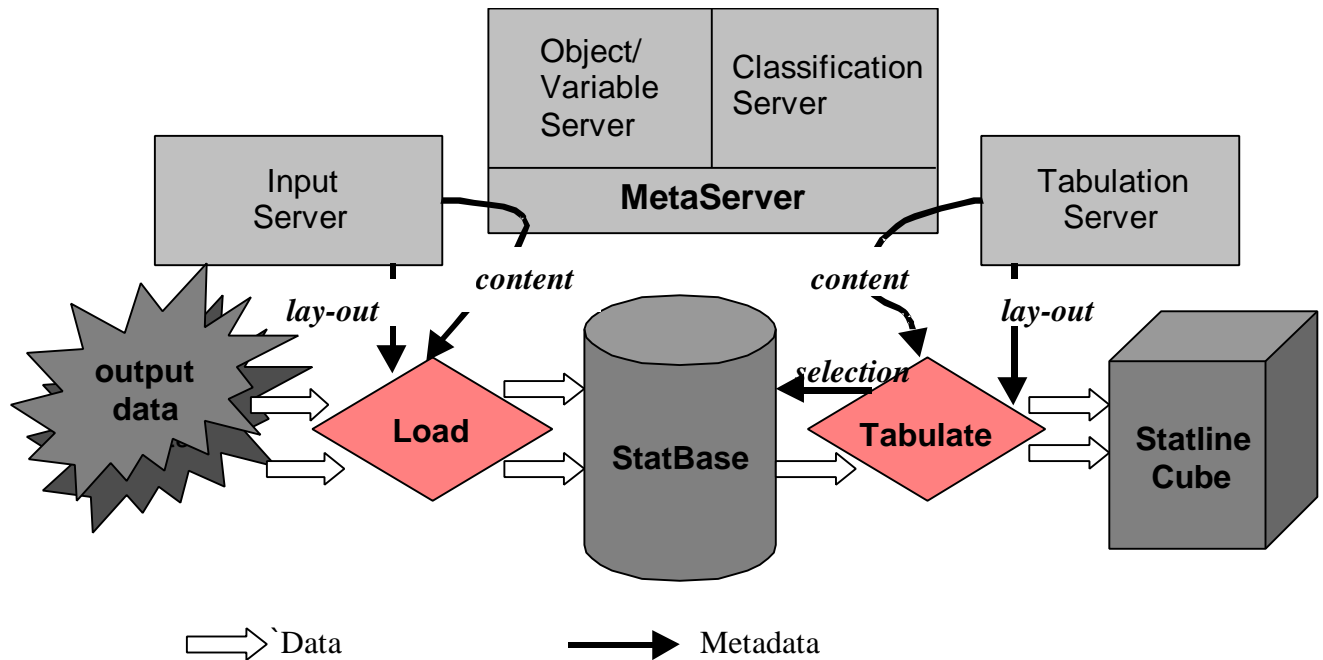
II.4 A tool for storage of meta: the MetaServer

20.   Obviously, from a coordination viewpoint, the MetaServer is the most interesting tool in the system. Not surprisingly, the structure of the MetaServer follows the structure of the model described in section II.2. Actually, there are two servers: a ClassificationServer ("MetClass") and a combined server for objects and count variables, called the VariableServer. Prototypes of both are presently being filled with data and tested in the process.

21.   The ClassificationServer is described in Ypma [1997]. It is a straightforward system for storage of classifications. One important aspect should be mentioned: classifications have *versions* that follow each other over time. From each version there is a number of variants, as initially not all statistics use the standard. This causes a coordination problem that will be discussed in the next section.

22.   The Object annex Variable server is more complex. It allows for objects (actually object types!) which may be linked to other objects as parent (inheritance of properties is not implemented yet, however). These objects may have several characteristics which may be count variables or classification variables. For the latter, one will have to refer explicitly to the classification server. The variables are described by their attributes in the database like a code, several (synonyms) names, definition, explanatory notes, etc.. It is also possible to define arithmetic relations between variables. This permits, for example, the calculation of total labor costs from its elements. Unlike classifications, (count) variables do not have variants. A "variant" will lead to another slightly different variable. In principle,

count variables can have *versions*, in the same way as can classifications. It is not yet clear how we will deal with versions of count variables.

23.   Both classifications and count variables have their "owners". The owners of the standards will have special responsibilities regarding statistical coordination. This aspect is further elaborated in section III. Both the ClassificationServer and the VariableServer are multilingual. Not only the software but also the data stored can contain more than one language. We concentrate on Dutch and English; sometimes also French and German will be available.

II.5 A tool for structuring data: the StatLine Datacube

24.   Not all aspects of the system can be dealt with. It is, however, interesting to pay some attention to the structuring of the data within StatLine. In StatBase, structure in this sense is not needed. It simply contains well-defined and well-described data ready for publication. Editors are needed to build the datacubes out of the StatBase building blocks in a way that fits the standard user needs best.

25.   Constructing a proper StatLine datacube proves to be rather difficult; indeed, it appears to be an art - the art of Cubism. With the data at hand in StatBase, to begin is simple. One selects data on a certain object-type. For a certain population of objects belonging to this object-type there will be information on several count variables and the total population will probably be divided into subpopulations using one ore more classification variables and classifications. An easy cube suggests itself. One axis will be used for the count variables and, for each classification used on the object, we have another axis. In practice this will often not suffice. More then five axes is not workable for the average user. So we will have to combine axes. Furthermore, data are often not available for every crossing between different classifications. Also, we might wish to combine different object-types, having one or more classifications in common, into one cube. It is obvious that this is not an easy job. What we do see, however, is that the datamodel is effective for the description and analysis of the problems at hand.

II.6 Tools for access: a thematic tree and a search engine

26.   Users look for data in StatLine. If the data are available they are "hidden" somewhere in a cube. StatLine has several instruments that help the user find his destination.

A thematic order of cubes

27.   At present StatLine contains about 300 cubes. One way of finding data is to search for the relevant cube through a thematic tree in which each cube has its place. In StatLine datacubes are grouped in themes and themes are grouped in main themes. How this thematic structure is set up will be described in section III. A priori, it is obvious that no thematic structure is perfect.

There are several safeguards that will prevent the user from looking in the wrong place: before he enters a theme or a cube he finds directly on his screen some more elaborate information on the contents; in the explanatory notes on the theme or the cube he will find pointers to related themes or cubes. All the other tools mentioned in this section will also be able to assist.

Offering predefined selections of data

28.   A Table Information Language (TIL) has been developed, enabling the storage of pre-defined selections (not the selected data itself) from StatLine datacubes. Such a selection can be applied anywhere: on explanatory notes within StatLine but also outside StatLine, for instance, on any HTML page. Recall of the selection will download the data from StatLine as they are at that moment. Furthermore, it is possible to define selections within TIL like "the last 5 years" ensuring the selection of the most recent data available.

The search engine

29.   Apart from going through the thematic structure, the user can type in questions such as "the depreciation in textile industry in 1995". The search engine tries to match the relevant words in these questions to any text within the cubes: labels, explanatory notes and explicitly added keywords (mostly synonyms). The user can opt for exact and for fuzzy matching.

30.   The search engine ranks the cubes hit by a score of success. Note that the meta used to match is relevant to one complete row or column in the cube and sometimes to the whole cube. The search engine ranks the cubes hit by probability of success. However, an important element of the search engine is that simultaneous hits for different parts of the questions in different dimensions of the cubes generate a higher score of probable success (textile industry in one dimension, 1995 in another and depreciation in yet another generate a high score.) The search engine gives some indication why this particular cube generated a hit.

31.   The search engine generates selections containing only the data as required in the search text from the cubes that scored. The user can check these selections one by one. If necessary he can change the selection within the cube he is working on.

II.7 Wider use of meta data

32.   Having described the actual role of meta data in the output process, we now indicate some further developments in the near future

Other parts of the statistical process: "globalizing" meta data

33.   As mentioned above, the use of the model for statistical data and metadata is not limited to the output side of the statistical process. It is

for a good reason that the MetaServer of the reference database is kept separate from the database for the output data itself. As much as possible we try to stimulate the use of the MetaServer for other processes. There is a micro database in use for data on individual enterprises (MicroLab). Here too, meta information is stored. No new system for meta information is built. The meta is drawn from the MetaServer around StatBase. This equality of meta is intended to become general policy.

34.   Recently we started to explore the possibilities to establish a link between BLAISE and the MetaServer. Eventually, we would like to go as far as to state that throughout the whole statistical process within the Bureau all meta should be drawn from the metaserver described here.

Other aspects of the statistical process: *"activating" meta data*

35.   The meta data discussed so far is still rather static. It describes statistical information, and does so in a most satisfactory way but we want more. A further step would be to *activate* meta data. Meta data will then describe operations on data in a way that can be understood by the processing system. The user can invoke the operations when needed.

III.   TOWARDS STATLINE 2002: FROM A COLLECTION OF PUBLICATIONS TO A COHERENT PICTURE OF SOCIETY

III.1 Introduction

36.   This section discusses the way the above-described tools are being utilized so that in 2002 the contents of StatLine can be said to provide a really coherent and comprehensive picture of society. But first of all, we must elaborate what is actually meant by such a picture. The state of ultimate coherence has a number of features that can be described as logical steps on the way to the ideal, where each step represents a certain level of ambition:

(i)   the first step is as trivial as it is important: establish *well-defined concepts*. It makes no sense to discuss the comparability between concepts if one does not know what they stand for.

(ii)   The second step relates to *uniform language*: if we know what our terms mean, we have to make sure that the same terms have the same meaning throughout StatLine and, conversely, that the same concepts are named with the same terms.(iii) The third step concerns *coordination*, which comes down to attuning (well-defined) concepts in such a way that they can be meaningfully related. Somewhat simplified, and with the StatLine datacube in mind, there are mainly two "directions" of relatability:

- *horizontally*: for two count variables (e.g. turnover and number of staff) to relate to each other, they have to refer to the same populations, and thus to the same object-type and the same classification(s);

- *Vertically*: for a count variable (e.g. number of staff) to be addible over the full range of a classification (e.g. for agriculture, manufacturing industry, trade, etc.), it must be defined equally for all classes of the classification. The coordination step results in a set of coordinated concepts, but it does not prohibit statistics from maintaining their own "nearby-concepts".

(iv) The fourth step therefore involves *standardization*. In order to protect users against a too broad and subtle - and hence confusing - assortment of nearby-concepts, these are eliminated as far as justified. Now that the concepts are clear and coordinated we can move to *data*:

(v) The fifth step consists of establishing *consistency* among *data* throughout StatLine. The most eye-catching expression of inconsistency occurs when, for the very same concept, StatLine shows different figures. More hidden forms of inconsistency are also possible.

(vi) The final step relates to the *presentation* of the data. They should be offered to the user in such a way that their relationships become maximally apparent. This step is implemented by first *structuring* each datacube so that it describes an area of interest and secondly by *ordering* all cubes in a thematic tree structure.

37. These are *logical* steps, each proceeding step representing a higher level of ambition, more difficult to achieve. In practice the order may be less rigid than suggested here, depending on specific circumstances.

III.2 Organization

38. The *Division for Presentation and Integration* maintains the output data warehouse StatLine. However, the statistical divisions who supply the data remain full owners of, and thus responsible for, their data: they "hire" rooms in the warehouse where they display *their* products.

39. A *Council of Editors*, in which all statistical divisions are represented, advises with respect to matters of common interest, such as editorial guidelines, the thematic tree structure and a centrally maintained list of synonyms. The Council supervises a number of editors, who have (limited) rights to change certain meta data (not the figures, of course) in StatLine.

40. Although not directly linked to StatLine, it is important to notice that the responsibility for coordination of concepts and consistency of data has recently been assigned to *Directors of Statistical Divisions*, on the understanding that the total field of statistical concepts has been distributed between them. For example, the Director of the Division for Labor Statistics has been assigned responsibility for (the coordination of) all concepts relating to labor, wherever in the Bureau such concepts may occur.

41.   Although the position of StatBase in the Statistics Netherlands organization is still under discussion, it is likely to be situated in the same Division as StatLine, i.e. the Presentation and Integration Division.

III.3 Policy

42.   How do we upgrade StatLine from the "collection of "electronized" paper publications", as it is predominantly now, to the coherent and comprehensive representation of society, easily accessible for external users *and* providing a solid basis for all Statistics Netherlands publications? In particular the two following statements can summarize the general policy:

- by maximally utilizing the *tools* described in the previous section and by minimizing the use of *rules*;
- by putting data suppliers, i.e. statistical divisions, in a position where they explicitly experience the need for and the benefits of coordination.

43.   The basic idea is that – as the past has taught - coherence cannot be achieved by a set of rules, issued by a centralized body "specialized" in coherence, but rather by providing attractive tools to statisticians who feel themselves responsible for coherence.

III.4 Implementation

44.   The following activities have been or are being underta ken:

- all publications, bought in StatLine during the past three years, were commented on - though not in depth - by StatLine editors. Producers of statistics complied with most of the suggestions, so that throughout StatLine there is now more uniformity in the way the meta is described and, to a certain extent, more uniform terminology and more precise concepts;

- Metaservers are being filled. This implies that statistical data sets have to be translated into the logic of the new datamodel described in Section II. This operation in itself requires more discipline in defining and distinguishing concepts and makes the relationships between the two more ransparent;

- StatLine editors screen the data in StatLine for inconsistencies and the concepts in the meta- servers for lack of coordination. They report anomalies to the appropriate Director(s), responsible for solutions to problems, according to the above-mentioned allocation of the total area;

- To enhance the power of the search engine "Hypersearch", a list of pure, quasi and fuzzy synonyms is being developed. This centrally maintained list connects concepts (via their terms) from different statistics. This

operation stimulates uniformity of terminology throughout StatLine as well as the elimination of redundant terms;

- In 1999 a new thematic 3-5 layer tree structure will be implemented, in order to guide thematic searchers to the appropriate datacube. Themes are delineated according to "areas of user interest", and therefore may trespass the borders of surveys and organizational units. Each theme is assigned a "theme-owner" (generally the survey manager supplying the majority of the data for the theme). The owner has a certain responsibility for a consistent and comprehensive description of his theme, regardless of the origin of the data. If concepts or data do not fit, he will negotiate with colleagues in order to attune these concepts or data;

- Next to "normal" themes like labor, education, health and environment, there are also some crosscutting "themes" according to a common classification variable. Outstanding examples are *regional* and *industry* breakdowns. The filling of these themes with count variables from different normal themes requires the use of the same classifications and therefore triggers standardization;

- Themes are described by one or more datacubes. A cube can be considered a manifestation and visualization of coherence. The challenge is to cluster all the data belonging to a certain theme in a minimum number of cubes, so that a user finds his data in one search-act *and* so that he can be sure that the data are relatable. The designing process of datacubes on the one hand confronts the surveying manager/theme-owner with the drawbacks of non-coordinated concepts and on the other hand rewards him with an elegant, compact and well-filled datacube in the case where the concepts fit in with each other.

45.   All actions listed here make use of the tools described in the previous section. It is obvious that these tools do not *generate* coherence and comprehension by themselves. Rather, they create the climate and the environment that provide statisticians with a stimulus to start coordinating action.

III.5 StatLine and Dissemination Policy

46.   Although the StatLine warehouse is not a publication in itself, it certainly is the spider in the Statistics Netherlands dissemination web. Indeed, as already mentioned, StatLine is the source from which *all* Statistics Netherlands publications are derived, either electronic or paper. Telephone information services of the statistical divisions use StatLine as their data source. This also goes for press releases.

47.   The information content of publications to be taken from the StatLine warehouse can be very flexible. Specific paper publications may cover one or more entire datacube, though this is not very likely in view of the large size of most cubes. Therefore, they will mostly relate to *selections* from one or

more datacubes. There are notably two publications which have a special position in that their information content is fully *identical* to that of StatLine. These are StatLine on CD-Rom and StatLine on the Internet. Whereas the price of the CD-Rom has recently been fixed at $1.000 annually (comprising 10 updates), the access to StatLine on the Internet data is free of charge.

REFERENCES

J.W. Altena and A.J. Willeboordse (1997): Matrixkunde of "The Art of Cubism" (Dutch only), Statistics Netherlands, Voorburg.

B. Sundgren (1992): Statistical Metainformation Systems, Statistics Sweden, Stockholm.

W.J. Keller and J.G. Bethlehem (1998): The Impact of EDI on Statistical Data Processing. Meeting on the Management of Information Technology, Geneva, February 1999.

W. Ypma (1997): Remarks on a Classification Server, Second Conference on Output Databases, Stockholm, November 1977.