

SEMINAIRE

E+; 3=! C

SEMINAR

STATISTICAL COMMISSION AND
 ECONOMIC COMMISSION FOR EUROPE



Distr.
 GENERAL

CONFERENCE OF EUROPEAN
 STATISTICIANS

CES/SEM.43/9
 7 April 2000

ENGLISH ONLY

Seminar on integrated statistical information
 systems and related matters (ISIS 2000)

(Riga, Latvia, 29-31 May 2000)

Topic I: Data warehousing and the development and use
 of statistical databases in a network environment

DATA WAREHOUSE IN A STATISTICAL OFFICE

Contributed paper

Submitted by Statistics Austria¹

I. INTRODUCTION

1. Statistical offices (National Statistical Institutes or NSIs) all over the world have faced many new demands, expectations and problems for some years now:

- The tasks they have to perform are increasing in complexity and scope.
- At the same time, manpower resources and funding are being frozen or cut back.
- "Data suppliers" would like to provide their information more simply and cheaply than hitherto.
- "Data clients" have a completely new means of seeking, collecting and using information: the main factor is no longer provision by others (e.g. provision by the staff of a statistical offices information service); instead, they are now used to collecting information interactively, online and on demand using appropriate search functions and processing it further on their own PCs. This constantly growing group of customers expects information providers to adapt to their way of handling information.
- The politicians', administrators' and the economy's needs in terms of up-to-date, high quality and internationally comparable statistics to help in decision-making processes is continually on the increase.
- Technical improvements in information technology (increasingly shorter innovation cycles) lead to major insecurity with regard to long-term investments. A product or technology opted for today may already be obsolete tomorrow.

1 Prepared by Günther Zettl.

2. Managing to deal with these needs and problems and being capable of reacting flexibly to future, completely unforeseeable developments, are herculean tasks. Bo Sundgren of Statistics Sweden wrote the following on this subject:

"It is a challenge for a modern statistical office to be responsive to expectations, demands and requirements from an ever more dynamic environment. Society itself, which is to be reflected by statistical data, is changing at an ever faster rate. This leads to needs for more variability, more flexibility, on the input side as well as on the output side of statistical information systems managed by statistical offices. In order to manage requirements for greater variability in the exchange of data with the external world, and in order to do this with the same or even less financial resources, a statistical office must consider system level actions. It is not enough just to do 'more of the same thing' or to 'run faster'. It is necessary to undertake more drastic redesign actions." [SUNDGREN 1996]

3. One-off activities are inadequate as system level actions - what is needed instead is a package of correlated organisational, statistical and technical measures. Since data are at the heart of the statistical production process and the computer is now the statistician's main tool, working out an overall strategic concept for the use of the computer (with special emphasis on metadata management; cf. [FROESCHL 1999a]) is an important matter.

4. There are various angles from which statistical production can be viewed. One of the simplest models is defined by a statistical office as a data processing system with two interfaces (fig. 1):

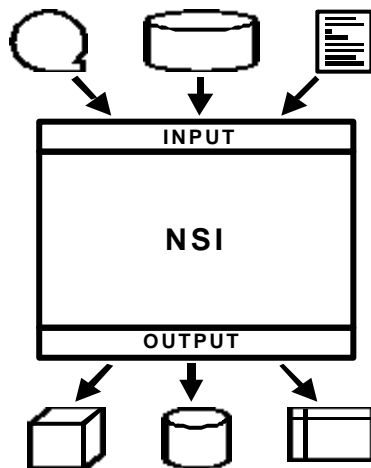


Fig. 1

- At the input end, raw data are entered into the "NSI system" by the "data suppliers" (respondents, existing data registers).
- At the output end, statistical results (object data and metadata at different levels of breakdown and in different forms of presentation) are passed on to "data users".

5. Many statistical offices are working on data processing projects aimed at modernising the "NSI system" including its interfaces and adapting it to new requirements:

- At the input end, the main focus is on reducing the burden for respondents (especially enterprises). One of the best known projects of this type is TELER which is being run by Statistics Netherlands. Statistics Austria, in close cooperation with an external software development firm, has started the SDSE project ("System zur Durchführung statistischer Erhebungen" - system for carrying

out statistical surveys), the central element of which is an "Electronic Questionnaire Management System" ("Elektronisches Fragebogen Management System" EFBMS).

- Inside NSIs, the systematic collection, administration and use of metadata is a basic challenge. A number of statistical offices have already started to build up integrated statistical metainformation systems (METIS).
- At the output end, printed publications are regarded as no longer adequate by "statistics clients". Here, efforts are concentrated on providing statistical results in electronic form. This includes projects involving the use of the internet for disseminating data, as well as the accelerated and standardised transfer of data to Eurostat using the STADIUM/STATEL software and the GESMES format. In Austria, the new Federal Statistics Law 2000 explicitly requires statistical results to be retrievable free of charge via the internet.

6. In discussions about the technical infrastructure of statistical offices and in the context of specific data processing projects (mainly in the output sector, but also within NSIs), there has been increasing mention of the concept of "data warehouse" recently. However, this term is sometimes used with a meaning which goes well beyond that of the original concept and therefore can lead to misunderstandings.

7. Therefore what is meant by a data warehouse (and related terms such as "data mart" and "OLAP") is described below. An attempt is also made to relate this concept to the statistical production process and to provide details of how data warehouse concepts and technologies can contribute towards meeting the challenges set for NSIs.

II. WHAT IS A DATA WAREHOUSE?

8. In the last few years, the term "data warehouse" has become fashionable in the computer industry:

- Hardware manufacturers love it because they can supply their customers with powerful computer systems for running a data warehouse.
- Software developers love it because they can sell expensive tools and applications (often costing millions of Austrian schillings) and do not have to compete with Microsoft (a situation which incidentally has changed in some sectors - data storage and OLAP - following the introduction of the MS SQL Server 7.0 and the accompanying OLAP services in the meantime).
- Consulting companies love it because their services are used by many companies which want to build a data warehouse.
- And authors of technical books love it because it is a wonderful subject for writing books and articles about. "The Data Warehousing Information Center" (<http://pwp.starnetinc.com/larryg>) currently (in December 1999) lists over 130 books, 70 White Papers and 100 articles accessible on the internet - which most probably only represent a fraction of the range actually on offer.

9. Of course there are a number of other definitions. Some examples can be found below:

- According to W.H. Inmon (often called the "Father of Data Warehousing") a data warehouse is "*a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process*" [INMON 1995].
- Ralph Kimball, with Inmon, probably one of the most famous "gurus" in the data warehouse field, defines a data warehouse as "*a copy of transaction data specifically structured for query and analysis*" [KIMBALL 1996].
- For Sean Kelly, a data warehouse is "*an enterprise architecture for pan-corporate data exploitation comprising standards, policies and an infrastructure which*

provide the basis for all decision-support applications" [KELLY 1997].

- Sam Anahory and Dennis Murray write: *"A data warehouse is the data (meta/fact/dimension/aggregation) and the process managers (load/warehouse/ query) that make information available, enabling people to make informed decisions" [ANAHORY/MURRAY 1997].*
- Barry Devlin describes a data warehouse as *"a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use in a business context" [DEVLIN 1997].*
- According to Stanford University, a data warehouse is *"a repository of integrated information, available for queries and analysis. Data and information are extracted from heterogeneous sources as they are generated.... This makes it much easier and more efficient to run queries over data that originally came from different sources" [STANFORD].*

10. Taken individually, none of these short definitions explain clearly what is meant by the term "data warehouse", but taken all together, they contain basic characteristics which will be described in slightly more detail below. However, it should be pointed out that, even if theoreticians agree on many features, there can be still differences in how the concept is understood in detail. For example, a data warehouse as understood by Inmon does not correspond to a Kimball warehouse in all aspects.

11. Originally, in the commercial environment computers were primarily used for supporting and automating business processes such as order-processing, invoicing, book-keeping, stock management etc. The aim was for these functions to run faster and more cheaply and for the company to be able to react more quickly to customers' demands. The main purpose of course was to derive advantages over competitors.

12. Computer systems in these areas of application are called OLTP (*Online Transaction Processing*) systems in data warehouse literature. They are optimised to allow fast response times for simple, pre-defined transactions which often consist in changes, additions or deletions of individual data records. By using normalised data modelling (preferably the third normal form, provided certain compromises do not have to be accepted for performance reasons) the aim is to ensure that the modification of a fact only has to be carried out on a single table line.

13. However, OLTP programs are not very suitable for providing information for analysis. Normally they allow certain reports to be issued but when further data are required individual programming by the IT division is necessary, if data are available at all (in a stock management system, for example, the current stock level can be determined, but the stock level of several months or a year ago or even earlier is no longer known).

14. Therefore, in view of their functionality and design, OLTP systems can hardly be used for analysis. To make up for this drawback, in the 1980s, it was proposed to extract data from them at regular intervals, provide them with a time stamp and store them in a system of their own: the data warehouse.

15. Since data mostly stem from several individually independent upstream systems, they may show a number of inconsistencies: e.g. different product numbers and descriptions in the programs for order-processing and stock management, non-uniform attributes for the same customers, when a firm is active in different business areas and uses more than one order-processing program, etc. Before the data are loaded in the data warehouse, therefore, they must undergo comprehensive integration, as well as structural and format standardisation (which sometimes represents up to 80 % of the total cost of establishing a data warehouse).

16. Unlike the more functional organisation of the OLTP systems, the placing of data in the data warehouse is oriented towards the main subjects for analysis (customers, products, supply companies etc.). Inmon calls this "subject-orientation".

17. The users of the warehouse should be able to find precisely the data they need for their work and carry out queries and analyses without the assistance of data processing experts. This requirement calls for special data modelling which is called "dimensional". To illustrate this data model, often a cube (fig. 2) is used, the edges of which are given the dimensions with their individual members (in the case of a warehouse for a chain of supermarkets, for example, product, outlet and time). Inside the cube, at the crossing of the different dimension members, there are numerical facts, e.g. turnover achieved on a particular day in a particular outlet for a particular product.

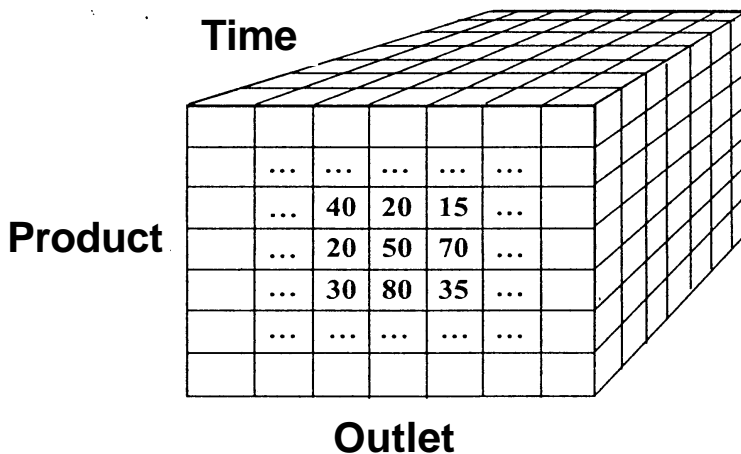


Fig. 2

18. The members of a dimension can be hierarchically broken down at several levels (e.g. product product group, outlet city district federal country, day month quarter year). They can also have attributes which might be of interest for analyses (the colour of the product, the selling space of an outlet etc.).

19. Queries and analyses of this type of "cube" (which of course may also have more than three dimensions) are called *Online Analytical Processing* or OLAP. OLAP client programs specialise in presenting the user with any required section through the cube in different arrangements of the dimensions (*slice and dice*). It is also possible to switch over from one hierarchy level to the elements below it and to navigate in the opposite direction (*drill down, drill up*).

20. OLAP cubes can be stored in a proprietary format in a multidimensional database: an OLAP server (MOLAP = multidimensional OLAP). Frequently the data are also located in a relational database (ROLAP = relational OLAP), in which case the so-called star scheme often comes into use. In a star scheme, each dimension with all its attributes and hierarchy levels is stored in a dimension table. The numerical values from inside the cube are stored in the central fact table together with the foreign keys of the dimension tables (fig. 3).

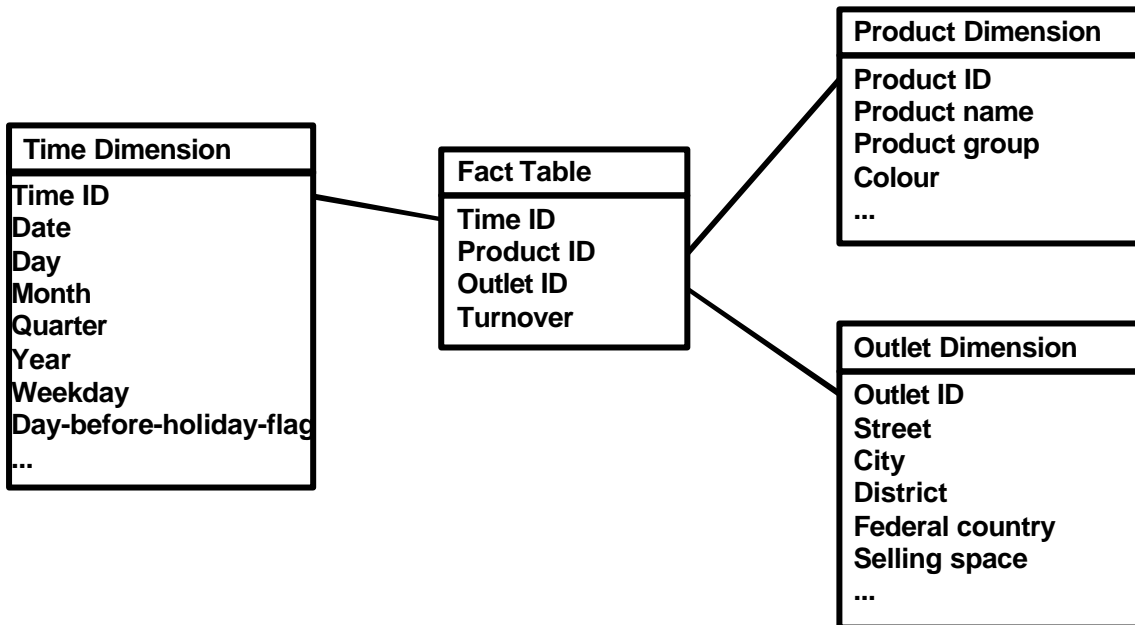


Fig. 3

21. An important characteristic of the star scheme is the denormalisation. For performance reasons (avoiding table joins) the attributes of objects which would be stored in separate tables in a normalised data model and would be referenced by primary/foreign key relationships, are entered in the dimension tables (in fig. 3, for example, the names of the town, district and federal country in the outlet dimension). This redundant data storage makes update operations difficult, which is why a data warehouse is normally read-only, or in Inmon-terminology, "non-volatile", for online users.

22. A data warehouse can contain massive quantities of data - on the one hand because of the implicit redundancy of the star scheme, and on the other hand because of the long periods for which data are stored. The level of detail (granularity) should also be as large as possible since otherwise potential possibilities for analysis are lost. In fig. 3, if we work on the basis of daily extraction of data and assume that in 500 outlets about half of 2000 products are sold at least once each day, the fact table will grow to just under half a billion records within 3 years!

23. To avoid accessing detailed data for every query, in a data warehouse frequently required aggregates are calculated in advance along the hierarchies of the dimension members and stored in their own tables. These advance aggregations speed up retrievals, but lead to an explosion in the amount of storage space required. In such circumstances, it is quite obvious that some warehouses have a size in the terabyte range.

24. Ralph Kimball is a keen defender of dimensional modelling. In his view, a data warehouse should consist of a number of star schemes, with thematically linked data cubes forming a data mart. Cross links between different marts develop through the use of uniform dimensions such as "customer" or "product", whereby consolidation and integration of the dimension data stemming from different advance systems take place in a staging area (which does not have to be relational, but can also consist of flat files).

25. Other authors such as W.H. Inmon, on the other hand, define a data warehouse as a company-wide normalised repository to which the end users can have direct access only in exceptional cases. From this central store, part quantities of data flow into divisional and functional data marts, which have a multidimensional structure. This

multi-layer architecture requires the development of a company-wide data model - a task whose complexity in practice is often made responsible for the failure of data warehouse projects.

26. In this connection, it should also be pointed out that the term "data mart" has not been clearly defined. Apart from the meanings already mentioned, it is also sometimes used to mean simply a "small warehouse".

27. A data warehouse contains not just data, but also all processes and programs required to extract data from upstream systems, to clean up, transform and load them in the warehouse, perform aggregations and to carry out queries/analyses are part of a warehouse (fig. 4). Basically, a distinction can be made between three subsystems:

- The input system in which the extraction and processing of source data and the loading of "cleaned" data in the warehouse takes place.
- The data-holding system which is responsible for storing and managing data (including aggregations and backup/archiving).
- The output system via which users access data stored in the warehouse via various tools (eg. report generators, OLAP client programs). This subsystem partly overlaps with the information factory (applications for further processing of data from the warehouse).

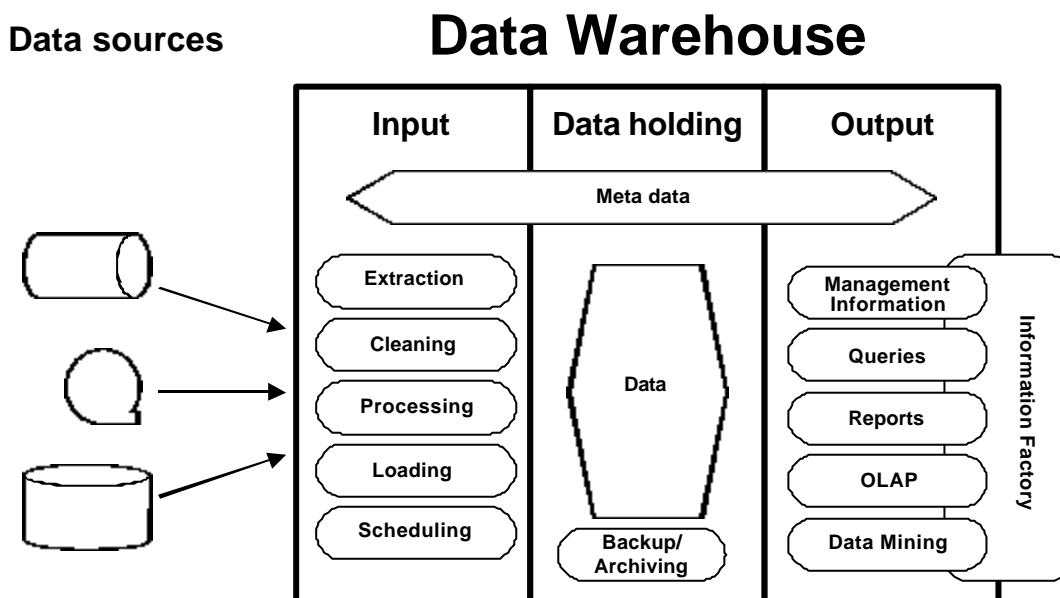


Fig. 4

28. In all three subsystems, there is also a need for metadata describing the stored and processed data. Ideally, there should be a central metadata basis used by all programs belonging to the warehouse. In practice, however, precisely the opposite is the case: the metadata of the tools which are used are incompatible with each other and have to be defined and administered independently of each other, which can result in a considerable amount of work in real operation. The Object Management Group (OMG) currently attempts to standardise metadata exchange (CWMI - Common Warehouse Metadata Interchange; information available at <http://www.omg.org/techprocess/meetings/schedule/CSMI-RFP.html>). Time will tell whether the proposals which have been presented are adopted and actually implemented by software manufacturers.²

² By the way: Eurostat consultants (Chris Nelson, Anders Tornqvist) from the company Dimension EDI are participating in the Common Warehouse Metamodel Specification. Their proposal for an "InformationSet" involves an extension of the warehouse concept that is important for statistical

29. In short, the following can be said:

- A data warehouse is a concept.
- A data warehouse is a process.
- A data warehouse must be constructed, in line with individual requirements.
- A data warehouse, however, is not an individual product, or off-the-shelf software. Of course there are a number of tools which cover some of the functions in a warehouse, and of course their manufacturers promise that all problems can be solved very quickly with these programs ("the 90-day-warehouse"). In practice, however, it is much more important to deal with conceptual, organisational, architectural and data-modelling questions. Whether any tools should be used, and if so which ones, is not important for the construction of a data warehouse until a relatively late stage.

III. THE STATISTICAL OUTPUT DATABASE AS DATA WAREHOUSE

30. The characteristics of a data warehouse mentioned in the previous section must seem familiar to any member of staff of a statistical office. For example, in statistics the multidimensional approach to facts has been around for some time now: in the form of cross-classified tables, which provide a two-dimensional depiction of data with several dimensions.

31. Other characteristics such as:

- very large quantities of data
- data stretching back over a very long period
- the need to validate, transform and integrate data
- hierarchical links between classification members
- the aggregation of detailed data
- the storage of these aggregates
- metadata which describe other data

are nothing new for an NSI - only the terminology used is different (classification criteria instead of dimensions, micro data instead of detailed data, time series instead of "time variant data") but not the underlying concepts.

32. These concordances between data warehousing and the statistical production process have so far been completely ignored in data warehouse literature, however. Historical articles locate the first data warehouses in the 1980s:

- *"Data Warehousing first emerged in this period between 1984 and 1988."* [DEVLIN 1997]
- *"The very first data warehouses were built in the USA in the mid 1980s by large corporations in the retail, banking and telecommunications industries. By and large, these early innovators were intent on integrating data that had become hopelessly fragmented across these complex organisations and the most common applications were (and still are) in the domain of marketing and sales."* [KELLY 1997]

33. But if we do not strictly limit the concept of the "data warehouse" to commercial enterprises and their business data, the first manifestations can be identified as early as in the 1970s - in the form of statistical output databases.

34. The statistical production of NSIs often has a structure which is termed

offices, namely the collection of raw data using electronic questionnaires.

"stovepipe" (cf. [PRIEST 1996]). Individual surveys, from the design of the questionnaires and the selection of respondents, and collection and processing of data up to the production of results tables and publications, are to a large extent implemented by different organisational units largely independently of each other. There is hardly any overall integration covering the entire "universe of surveys". Each "stovepipe" can be regarded as an independent statistical information system, which leads to a lot of problems (lack of overview over the entire system; unplanned redundancy with resulting higher maintenance costs and the danger of inconsistencies; disharmonies and discrepancies with regard to statistical concepts, definitions, variables, classifications, results etc.; no standardisation of data holding and processing; little reuse of software).

35. A statistical output database (fig. 5) combines object and metadata from separate pre-systems - in which it is not OLTP programs that are involved as in a typical warehouse but surveys - in a central application which aims at simple retrievability of data by the users. This does not allow the problems mentioned to be eliminated in retrospect, but for the "statistics client", the existence of an output database compared with a pure "stovepipe" organisation brings with it considerable advantages. Ideally, all information published by an NSI in publications, press bulletins, WWW pages etc. should also be contained in the output database in greater detail or be derivable from it.

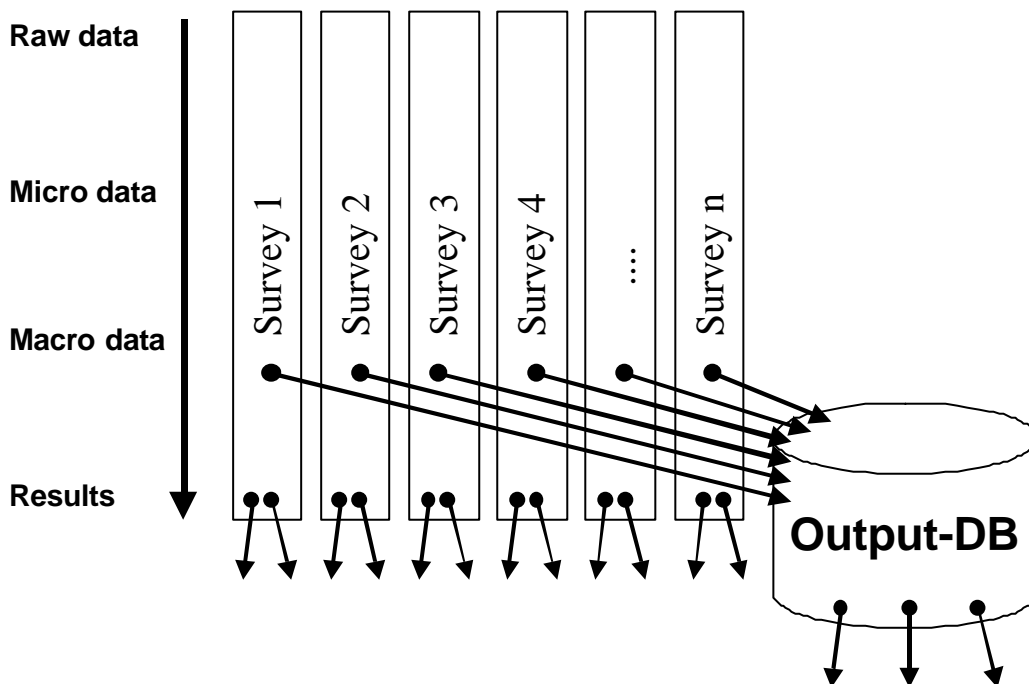


Fig. 5

36. At Statistics Austria, the output database ISIS ("Integrated Statistical Information System", also known internationally as LASD - "Large Scale Statistical Data System") was developed as early as 1972/73 - many years before the expressions "data warehouse" and "OLAP" were invented. Nonetheless, ISIS can be described as a MOLAP server in current terminology:

- It consists of over 4000 multidimensional cubes in a proprietary format, with a maximum of 7 dimensions possible per cube.
- In all there are several hundred dimensions which can be structured hierarchically.
- Some aggregations of hierarchical dimensions are calculated in advance and stored when the data are loaded whereas others are calculated "on the fly" when they are

retrieved.

- With the help of a powerful query language, the user can retrieve results with any dimensions and at different hierarchical levels (slice, dice, drill down, drill up). There are also numerous mathematical/statistical functions available.

37. In addition to the object data, ISIS also contains metadata: in order to find a specific cube (e.g. via a hierarchically structured list of subjects or via full text search) and to find information about the data contained in a cube (data source, breaks in time series, etc.).

38. ISIS runs on Statistics Austria's IBM main frame and consists of about 800 assembler and PL/1 modules (including various administrative programs for the database administrator). In view of its early origin (E.F. Codd had only just published the theoretical deliberations which were to serve as a basis for relational database systems) ISIS had to be developed down to the finest details at Statistics Austria. If we began with the implementation today, we would probably use object-orientated programming languages, design an n-tier-architecture using DCOM or CORBA and include for data storage a relational database and possibly also a commercial OLAP-server like Microsoft OLAP Services or Hyperion Essbase. But even if ISIS is no longer at the latest level of IT, it is still state of the art in its applied concepts and ideas!³

39. There is no denying that Statistics Austria was subjected to three points of criticism regarding ISIS in the past few years:

- i) The content is not always completely up-to-date.
- ii) The user interface no longer meets modern expectations.
- iii) The query language is difficult to learn and is quickly forgotten again if it is not used regularly.

40. The first point of criticism is connected with the ranking of the output database within the production of statistics. For some divisions, the provision of data for ISIS would appear to be a necessary evil that is not dealt with until other work (production of publications in different formats) has been completed. With appropriate management decisions at Statistics Austria, which has a new structure as from 1.1.2000, and with organisational arrangements, this problem should be solved relatively easily.

41. As far as the second point of criticism is concerned, this is a challenge for the informatics division. At the moment, work is being done on a graphical user interface which allows ISIS queries from any Java-compatible WWW-browser, but can also be started as an independent application. Since a knowledge of the proprietary ISIS query language is no longer necessary when using this new client software, the third point of criticism is dealt with at the same time.

42. Apart from the fact that a statistical output database contains no detailed data, it has all the main characteristics of a data warehouse. However, if one mentions to data warehouse experts working in a commercial environment on the basis of publications by Inmon to Kimball, that large statistical databases contain thousands of multi-dimensional cubes and hundreds of different dimensions, one is normally confronted with reactions such as:

- *„I would also be wondering who in the world could possibly mentally manage 113 dimensions in one multidimensional model. People have difficulty conceptually managing much more than seven or so dimensions in one place, even though the tools*

³ And there were no Y2K-problems!

allow more."

- „I would suggest re-visiting your design, especially if it has 100+ dimensions. I would also re-consider your fact-table design, especially if you have 100+ of those."
- „I too spent twenty years working at a National Statistical Agency (Statistics Canada) and have followed this and other threads discussing the perfect dimensional model of 6-12 dimensions and a couple of fact tables with a great deal of interest. Over the past few years I have presented the Canadian Census model at various local and international venues, and have been told by the 'experts' that the model, composed of hundreds of dimensions was poor design and planning, and either could not work or would be impossible to manage."

(All these quotations are taken from contributions to the „Data Warehouse Mailing List"⁴ from the end of November/beginning of December 1999 in response to a mail which was written by a member of a Statistical Office staff - whichever office it was, unfortunately was impossible to identify from the e-mail address).

43. Where do these many dimensions in a statistical output database come from?

44. The reason for the "dimensional explosion" lies in the function to be performed by such a database: it is supposed to provide the results of statistical production processes in multi-dimensional form for queries by end users. Statistical information is information not about individuals but about collectives, in other words an NSI can only publish aggregated data - for legal reasons for a start. This is why an output database contains no data at the most detailed level.

45. If we take for example the Census (including the accompanying full survey of housing) which is held once every ten years, it becomes immediately clear that the star scheme is not particularly appropriate for it. If advanced concepts of dimensional modelling such as "demographic mini-dimensions" (cf. [KIMBALL 1996]) are not applied, a star with only two dimensions is probably obtained: namely "Person" (with numerous attributes such as "sex", "age", "marital status", "nationality", "number of children" etc.) and "housing" (with a regional hierarchy and also a number of attributes). A time dimension does not exist because all person-related data have to be rendered anonymous so that it is not possible to retrieve the Census data of e.g. 1981 and 1991 for the same individual.

46. On the basis of the number of data records, it would also be a very unbalanced star scheme. Normally dimension tables contain relatively few and fact tables very many records (cf. the example mentioned previously of a data cube for a supermarket chain with 500 entries in the outlet dimension, 2000 in the product dimension and just over 1000 - three years on the basis of daily data extraction - in the time dimension: when an average of 50 % of the products per day and outlet land in the customers' shopping baskets, the fact table shows just under half a billion lines after three years). By comparison, the "Census star" in Austria would have about 8 million data records in the person dimension (in the USA it would be over 200 million even!) and 3.5 million in the housing dimension; the facts table on the other hand would probably not exceed ten million records.

47. A statistical output database, as mentioned, contains no detailed data but many relatively small data cubes resulting from summations using individual variables of the survey. For example, a cube could depict the fact "Number of Persons" and the dimensions "Time" (which is available again for aggregates but not for the finest level of detail), "Region" (a hierarchy with "Municipality", "District" and "Federal country" levels), "Sex", "Age" (with hierarchical age categories) and "Marital

⁴ Subscription possible at <http://www.datawarehousing.com>

Status"; another cube could depict the dimensions of "Time", "Region", "Nationality", "Number of Children" and "Age", etc. This means that attributes at the level of detailed data become dimensions of one or more aggregated data cubes. This is why, in an extensive statistical output database, a "dimensional explosion" occurs, a phenomenon which I have so far not seen described in any data warehouse literature I know of.

IV. THE NSI AS DATA WAREHOUSE

48. As shown in the previous section, we are fully justified in calling a statistical output database a data warehouse. We could now go one step further and call the entire statistical office a warehouse. For this we should look at the „NSI System“ from Fig. 1 a bit more closely.

49. To solve the problems resulting from „stovepipe“ organisation, work is done in many statistical offices on concepts and projects for managing object and metadata and on developing integrated statistical meta information systems. The aim of these activities is horizontal (in other words multi-survey) and vertical (going beyond the stages of statistical production) integration of statistical information systems into a universal information infrastructure.

50. The general aims are:

- to create as extensive, flexible, open, simple and user-friendly access as possible to the object and metadata of relevance to them for both NSI internal and external „statistics users“;
- to plan redundancy and avoid inconsistencies;
- to achieve planned collection, storage and (multiple) use of metadata (which means that they have to be standardised and harmonised);
- to establish norms for data-holding in general and for interfaces between the software products used for producing statistics;
- to provide support to users with general-use tools, in other words, tools which are not tailored to a single survey only, in performing their tasks in the production and use of statistics;
- to enforce global solutions, in other words solutions covering the entire statistical office, instead of insular solutions or double and multiple developments;
- and finally, always to take into account the diverging, and in some cases unknown or unpredictable needs of different user groups.

51. These demands require the setting up of an NSI-wide information system covering all surveys and supporting the entire statistical production process - from preparations for a survey to the dissemination of results - with appropriate tools.

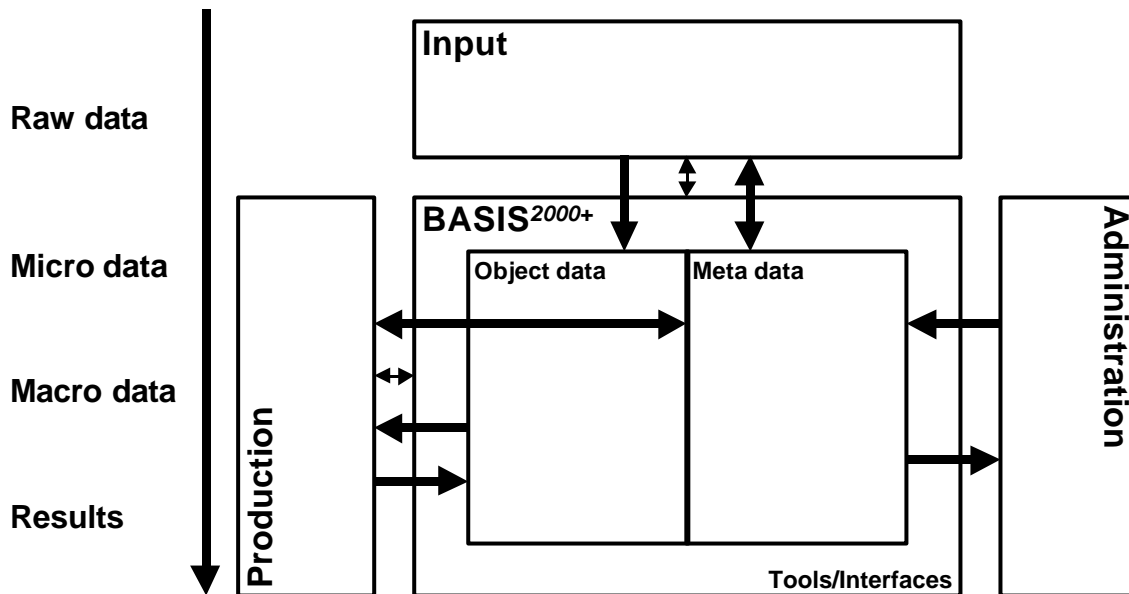


Fig. 6

52. Fig. 6 shows the **BASIS²⁰⁰⁰⁺** (metadata **b**ased **s**tatistical **i**nformation **s**ystem) concept developed by Statistics Austria together with the links and data flows to other systems. **BASIS²⁰⁰⁰⁺** consists of three components:

- The object data component contains statistical data with different levels of aggregation. The first stratum is formed by the checked and corrected micro data of all the surveys conducted by Statistics Austria. In the course of statistical production, micro data are aggregated into macro data, which also form part of the object data component and in many cases serve as a basis for further evaluations. The highest level of aggregation is statistical information in the form of tables and graphs; the aim must be to have all the "information objects" produced by Statistics Austria accessible in **BASIS²⁰⁰⁰⁺**.
- The metadata component is at the heart of **BASIS²⁰⁰⁰⁺**. Here the content of the object data component must be documented in order to allow both physical access to object data and the interpretation of their content. With its information about classifications, surveys, statistical concepts, variables, data holdings, publications etc., the metadata component forms an extensive reference database which opens up Statistics Austria's statistical information to both internal and external users.
- The information stored in the object and metadata components is accessed via user and program interfaces. These should be regarded as parts of the third **BASIS²⁰⁰⁰⁺**-component together with standardised data formats and generalised tools which are used at different stages of the statistical production process.

53. The life cycle of a statistical survey begins with the "observation" stage which covers all activities connected with preparing and planning the survey and collecting data. These take place in the "Input" system in which metadata already stored in **BASIS²⁰⁰⁰⁺** are accessed and new metadata are added. As soon as this information is available in the metadata component, tools from the tools/interface stratum of **BASIS²⁰⁰⁰⁺** can make use of them.

54. In the "Input" system also the second sub-part of a survey ("preparation", in other words, collection, checking and correction of raw data) takes place. Here, too, apart from individually developed software, general-use tools can be employed and the content of the central metadata component is accessed. Finally, the micro data

regarded as correct are loaded into **BASIS²⁰⁰⁰⁺** in a standardised format. The micro data are aggregated into macro data at the next stage of statistical production. This is done partly by an automatic process within **BASIS²⁰⁰⁰⁺** and partly in the "Production" system, whereby the macro data holdings produced there are again loaded in the object data component and supplements/updates are carried out in the metadata component.

55. This division of tasks also applies to the "use" process: the micro and macro data provided in **BASIS²⁰⁰⁰⁺** (including the accompanying descriptive information) serve as a basis for all analyses. These are either processed with general-use tools or exported into the "Production" system (where for example analyses are carried out with SAS or individual software). Information objects produced in the "Production" system - finished tables, graphs, documents - are stored in **BASIS²⁰⁰⁰⁺** in standardised formats and documented in the metadata component. The "use" process also includes the search for and retrieval of statistical information by external "clients"; this is done exclusively in **BASIS²⁰⁰⁰⁺**.

56. Fig. 6 also shows the "Administration" system which comprises all applications which are not or only indirectly connected with statistical production (e.g. a staff information system). Data flows occur between "Administration" and the metadata component of **BASIS²⁰⁰⁰⁺**, for example when the current telephone number of the person responsible for a survey is the subject of an inquiry.

57. It should be emphasised that **BASIS²⁰⁰⁰⁺** is not a single, monolithic application. Instead, it consists of minor sub-systems which are integrated on the basis of the jointly used standardised object and metadata. **BASIS²⁰⁰⁰⁺** is primarily a general concept - a vision. It provides a framework which allows implementation to start in sub-sectors, prototypes to be developed with results which can be used for practical applications and experience to be collected as quickly as possible - which in turn allow the overall concept to be refined and adapted in a feedback process.

58. This was a short review of **BASIS²⁰⁰⁰⁺**. Are we now justified in speaking of a data warehouse?

59. I don't think so. Of course there are a number of parallels to the data warehouse concepts, processes and features (especially if we define a warehouse not as a collection of star schemes but as a company-wide, multi-tiered integrated data repository), but the scope and extent of an architecture such as **BASIS²⁰⁰⁰⁺** goes far beyond the meaning that is associated with the term "data warehouse" by 95 % of IT experts.

60. The concept of "metadata" alone has to be interpreted on a very much broader basis in a statistical environment. Statistical data are always a combination of object and metadata, whereby the latter are both produced in the statistical production process and re-enter the process as input in other work stages.

61. Statistical classifications, for example, can be both metadata (texts being allocated to codes) and independent complex "information objects" available in different versions, whose elements may have links with elements of other versions and classifications and to which metadata (e.g. a list of technical terms allocated to classification members) belong as well. Accordingly, a classification database for administering classifications and the accompanying metadata is a central feature of the metadata component of **BASIS²⁰⁰⁰⁺**, whereas in a data warehouse in a commercial environment, such an application is not required.

62. We get even further away from the typical warehouse when we begin to place the emphasis, when considering statistical meta information systems, not so much on descriptive metadata oriented to human users but instead concentrate on procedural aspects (active, "embedded" metadata, the meta information system as a "workbench";

cf. for example [BETHLEHEM et al. 1999] and [FROESCHL 1999b]).

63. Since even describing a statistical output database as data warehouse can lead to misunderstandings in discussions with warehouse experts from the commercial sector, it would appear sensible not to call NSI-wide (meta) information systems "data warehouses" - we can thus save ourselves a lot of explaining.

Literature

[ANAHORY/MURRAY 1997] Sam Anahory/Dennis Murray, *Data Warehousing in the Real World*, Publishers: Addison-Wesley, ISBN 0-201-17519-3

[BETHLEHEM et al. 1999] Jelke Bethlehem, Jean-Pierre Kent, Ad Willeboordse and Winfried Ypma, „On the Use of Metadata in Statistical Data Processing“, Report for the „UN/ECE Work Session on Statistical Metadata“ in Geneva, 22 to 24 September 1999

[DEVLIN 1997] Barry Devlin, *Data Warehouse: from architecture to implementation*, Publishers: Addison-Wesley, ISBN 0-201-96425-2

[FROESCHL 1999a] Karl A. Froeschl, „Metadata Management in Official Statistics - An IT-based Methodology Approach“, in *Austrian Journal of Statistics*, Vol. 28 1999 Number 2

[FROESCHL 1999b] Karl A. Froeschl, „On Standards of Formal Communication in Statistics“, Report for the „UN/ECE Work Session on Statistical Metadata“ in Geneva, 22 to 24 September 1999

[INMON 1995] W.H. Inmon, „What is a Data Warehouse?“, published in the World Wide Web at http://www.cait.wustl.edu/cait/papers/prism/voll_no1

[KELLY 1997] Sean Kelly, *Data Warehousing in Action*, Publishers: John Wiley & Sons, ISBN 0-471-96640-1

[KIMBALL 1996] Ralph Kimball, *The Data Warehouse Toolkit*, Publishers: John Wiley & Sons, ISBN 0-471-15337-0

[PRIEST 1996] G. Priest, „Issues of Meta Information and Integration“, Report for the „UN/ECE Work Session on Registers and Administrative Records in Social and Demographic Statistics“ in Geneva, 11 to 13 November 1996

[STANFORD] The quotation ascribed to Stanford University which was published in the World Wide Web at <http://www.datawarehousing.com>

[SUNDGREN 1996] Bo Sundgren, „Making Statistical Data More Available“, in *International Statistical Review* (1996)