NATIONS UNIES   ОБЪЕДИНЕННЫЕ НАЦИИ   UNITED NATIONS

| COMMISSION ECONOMIQUE POUR L'EUROPE | ЭКОНОМИЧЕСКАЯ КОМИССИЯ ДЛЯ ЕВРОПЫ | ECONOMIC COMMISSION FOR EUROPE |

| SEMINAIRE | СЕМИНАР | SEMINAR |

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

EXPERIENCES WITH DATA ARCHITECTURE AND WWW-BASED DATA DISSEMINATION AT STATISTICS FINLAND

Contributed paper

Submitted by Statistics Finland[1]

I. THE DATA ARCHITECTURE

I.1 Developing the data architecture

1.   During the period 1994-1997, the data architecture project at Statistics Finland concentrated on developing and building, what we called, the customer-oriented implementation model for the data architecture. One of the aims was to develop the data architecture based on the Unified File System and the Classification Database, which had been built earlier into the mainframe environment, and a suitable multidimensional database. The development of metadata processing had a high priority. The implementation environment would now be PC/Unix (Powerbuilder, Sybase SQLServer). The planning of the implementation model was based on the principle that the data architecture can be built and put into use step-by-step. Following the introduction stage of each step the data architecture model can be revised in the implementation environment to conform to the general development of information technology. However, everything did not go as planned.

2.   The planned and partially carried out implementation model consists of three parts: the file part, the statistical product part and the metadata or description part. The file part is associated with the data manipulation stage before the data are transferred to the final product stage, i.e. to statistical products. The implementation model defines the framework according to which the data of a statistical system transforms from basic data to a statistical product, i.e. the path of the statistical system's data from one step of the three-level hierarchy of files (separate files, unified files and multidimensional databases) to a statistical product. The goal was that most of the statistical products would be made of

---

standardised and well described files (unified files and multidimensional databases) and the files from the lower level of the hierarchy (separate files) would be used in this context as little as possible.

ON LINE DATABASES

MESSAGE BASED STATISTICAL SERVICES

TAILORED INFORMATION SERVICES

INTERNET SERVICES

CD AND DISKETTE PRODUCTS

FAX SERVICES

PAPER PUBLICATIONS

STATISTICAL PRODUCTS AND SERVICES

SEPARATE FILES

UNIFIED FILES

ARCHIVES

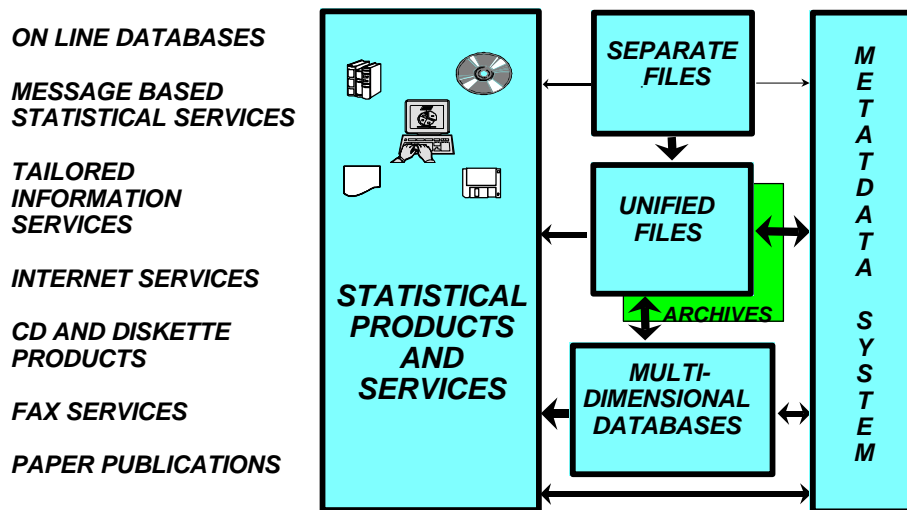**MULTI-DIMENSIONAL DATABASES**

METADATA SYSTEM

FIGURE 1: The customer-oriented implementation model for the data architecture

3.   The project succeeded in building the Unified File System with a description database which contains the technical and general descriptions of U-files and their fields and has connections to the Classification database. The classification Database can also be used independently.

4.   We have used mainframe based on-line output databases since the 1980s. These databases have been quite advanced and functional so that users, often working with "dumb terminals" could process the data using the services of these databases. During the period 1995-1997, the project tried to build a new version of our public online database into the PC/Unix environment. This did not succeed as well as we had hoped. The biggest problem was lack of resources and so at the beginning of 1998, it was decided to halt the development of this work.

I.2 The Data architecture now

5.   In 1998, Statistics Finland began searching for a new data warehouse solution for storing aggregate data in table format for internal use and for disseminating this data, or parts of it, using WWW-technology and CD-ROM discs. A project was launched to propose a production model and it came to the conclusion that the StatLine-output system, made by Statistics Netherlands was the tool to use in implementing this production model. The implementation process is described in the second part of this paper.

6.   At the same time, as we chose to put into use the StatLine-system, we also started to develop the Concept Database. Both of these systems fit well into the earlier defined implementation model for the data architecture.

7.   Figure 2 shows a simplified production model that we are using now. This is not, however, the final model. We think that interfaces between different systems are important. One direction for developing the data architecture is to build feasible interfaces and allow statisticians to use the tools that they know and are suitable to the problems at hand.
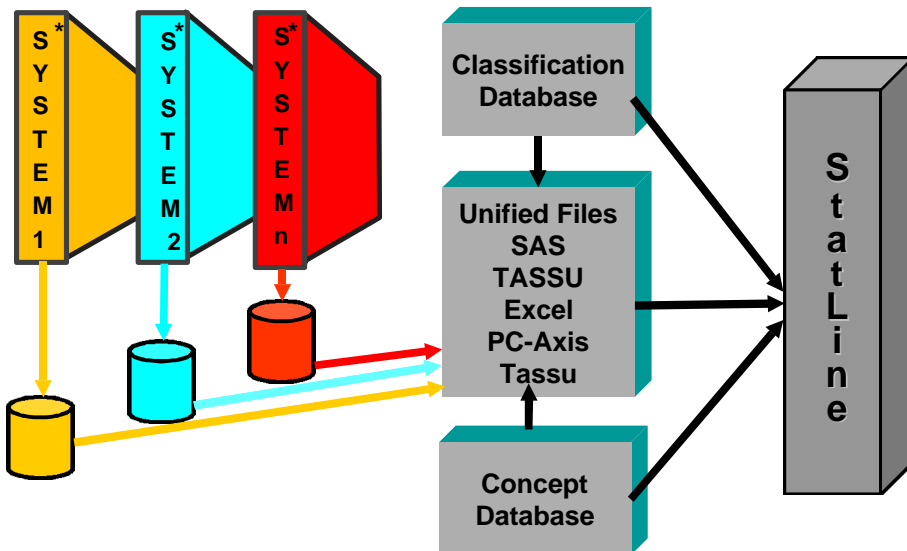
* SYSTEM = Statistical data system = Survey

FIGURE 2: A simple production model

8.   The unified files (U-files) are files belonging to the unified file system built at Statistics Finland. They can be sequential MVS files, Sybase tables or SAS files. The fields of U-files are classified as background variables, quantity variables or index variables according to their later purpose of use. If a field of a U-file is a classification variable, it is possible to build a connection from it to the Classification Database. If a field of a U-file is a quantity variable, it will be possible to build a connection from it to the Concepts Database. U-files can be used in normal data production, but most frequently they are used after the maintenance stage when the data is ready for output.

9.   A description of a U-file consists of the following elements:

♦ general description of a file (text);
♦ technical description of a file;
♦ general description of the fields (variables);
♦ technical description of the fields;
♦ connections to the Classification Database and the database for statistical concepts;
♦ keywords for search function.

10.   The Classification Database is the general location to store classifications. The database contains the following information:
♦ classification codes and texts;
♦ definitions of the items (explanatory notes);
♦ index entries of the items (title words);
♦ keys between classifications (correspondence tables);
♦ rules of aggregations.

11.   There can be several versions of a classification and there can be different versions for different years when necessary. A user can, if he/she so wishes, create his/her own version of a classification and store it in the database. There are also text versions for abbreviations and for different languages (Finnish, Swedish, English). It is also possible to export classifications to other formats (ASCII, Excel, HTML, CLASET).

12.   The Concept Database is the latest addition to our Metadata system. In summer 1998, a project was launched to construct a database that would contain all the statistical concepts and definitions used in our statistical production. Now that the

technical solution is completed it is time to market the database to operational units. The key persons are to be found among the producers of the Internet- and database services. These people will benefit the most from an "electronic warehouse" as their work consists largely of updating and combining of statistical data from different units with various concepts and classifications.

13.   In the database, you can find the statistical concepts and definitions and their different versions:
- short and long definitions of concepts (explanatory notes);
- text versions for different languages (Finnish, Swedish, English);
- different text versions of concepts and definitions used in different operational units and by other national and international producers of statistics.

14.   Furthermore, the database contains references and comparative information for the concepts. Relevant classifications, related concepts and index entries are defined for all the concepts. Comparison and analysis of the similarities and differences between the concepts, definitions and versions of concepts can also be included in the database.

15.   At the moment, the metadata system of Statistics Finland consists of the following elements:

♦ The description database of the Unified File System,

♦ The Classification Database,

♦ The Concepts Database,

♦ The System Register,

♦ The metadata parts of the StatLine system.

The first three of the above-mentioned elements of the metadata system play an important role when producing multidimensional tables for the StatLine database.

16.   At the moment, it is possible and recommended to produce StatLine tables with required metadata direct from a U-file. However, it is also possible to import a PC-Axis-file, an Excel table, a SAS-file or a TASSU table (TASSU is our own home-made tabulating system) to StatLine using conversion programs, but then you have to bring metadata to StatLine-system from other sources.

II.   EXEPRIENCES WITH WWW-BASED DATA DISSEMINATION – THE STATFIN ON-LINE SERVICE

II.1 Summary

17.   This part of the paper describes Statistics Finland's experiences in introducing a WWW-based on-line dissemination database - the StatFin on-line service. Statistics Finland has had on-line data dissemination databases for years, but these systems are no longer able to meet the needs of the users, nor are they seen as dissemination databases for the whole office. The change from centrally coordinated on-line output to a distributed, but still centrally administrated WWW-based on-line service was not an easy one - especially when it was to be made within one year. Still the project responsible for this change seems to have succeeded as the feedback from the users is highly positive and the service already contains over 30 million data elements within over 100 tables.

II.2 Background

18.   Statistics Finland has a long history of providing on-line output database services. We have used mainframe-based on-line output databases since the 1980s and

made them accessible through the Internet. This happened long before the WWW-
environment made its breakthrough in the world of Internet. These databases were
quite advanced and were very functional so that the users, often working with "dumb
terminals" could process the data using the services of these databases.

19.   The data architecture of Statistics Finland describes the way the data should
flow from production databases to output databases (Saijets 1999). All seemed to be
well in theory, but in practice there were a vast number of separate production
systems using their own data sources and their own dissemination channels. This
situation has been called the stovepipe-model (Keller 1998).

20.   Since then, a lot has changed both in the data processing facilities of users
and in the use of networks, the Internet. At the beginning of the 1990s, the
Internet became a phenomenon known by almost everyone, and the traditional users,
universities and the science community, faced new challenges when the WWW-technology
brought millions of new users to the services of the network. Statistics Finland was
also faced with the pressure introduced by this evolution, as its on-line databases,
based on mainframe solutions and direct terminal access were suddenly seen as hard-
to-use "dinosaurs".

II.3 Definition of policy

21.   In 1998, Statistics Finland began searching for a new data warehouse solution
for storing aggregate data in multidimensional table format for internal use and for
disseminating this data, or parts of it, using WWW-technology and CD-ROM discs. The
selection process was a long and difficult one. There were many options to choose
from, but none of them seemed to cover all the needs the office had. The main
choices were to:

a. Continue with our current situation (stovepipes), to further develop the output
   side and enhance the co-ordination by strict rules.
b. Select one of our data-streams and tools related to it as an office-wide standard
   (one of the candidates was PC-Axis-based distribution, which has been used since
   1992).
c. Look for other solutions and replace our current production model and tools with
   them.
d. Look for a tool strong enough to solve the most obvious (output) problems and to
   integrate it to our metadata and production systems, thus increasing the degree of
   integration.

22.   Owing to the diversity of the options, a project was launched to evaluate the
options and to make a proposal for a production model. This project was called the
StatFin2000-policy definition project. The project did evaluate the different
options and systems and proposed a production model where a central data-warehouse
acted as a storage for aggregate data in table format. The project also came to the
conclusion that the StatLine-output system, made by Statistics Netherlands was the
tool to use in implementing this production model.

23.   The decision was not an easy one - nor was it accepted by all parts of the
organisation. This imposed heavy pressure on the success of implementation of the
selected production model and the tools.

II.4 The goals of the StatFin2000-implementation project

24.   The StatLine system is made in the Netherlands. It was, therefore, necessary
to implement it in the production environment of Statistics Finland. There was a lot
of work to be done and the schedule was tight (one year). The main objectives set
for the implementation of the project were to:

a. localise the program suite (the StatLine Suite) for use in Statistics Finland
   (definition of concepts, translating the programs and manuals);
b. purchase the software and hardware required for using the system;

c. adapt the other production tools used in Statistics Finland to seamlessly integrate into the StatLine system;

d. plan the best practices (using pilot-systems) to use the data warehouse and information service (StatLine system) in different cases and to prepare the instructions and manuals needed;

e. plan the structure of the internal data warehouse and the WWW-service and to help and guide the statistical departments in providing the content to the system;

f. arrange the necessary technical support, courses and training;

g. guide and supervise the office-wide introduction and implementation of the system;

h. organize and find/appoint resources for administrating the system after the project has ended.

II.5 Resources

25.   Due to the importance of the success and the tight schedule, the project was financed very well. For its one-year duration, it received a budget of 1.5 million Finnish Marks (approx. 254,000 Euros). This enabled the project to recruit enough personnel with a wide scale of expertise that the complexity of the work required. The preceding project (the StatFin2000-policy definition project) also obtained 300,000 Finnish Marks (approx. 50,500 Euros) to buy the server-machines needed for the implementation.

26.   In addition to the financial resources, the project also received the full support from the top-level management of the office, which proved to be even more important than the money.

27.   The project's consumption of resources was quite stable. During the first four months there were five persons in the working group (some of whom worked for the project on a part-time basis only) plus the project manager. Later, one full-time staff member was hired to the project which subsequently increased the consumption of resources accordingly. At the beginning of October another staff member was added to the working group.

28.   As the project introduced an office-wide change, it also incurred a lot of costs for other organisation units, but at the time of writing this report (December 1999), exact data on those costs are not yet available. As a conclusion, it could be stated that the resources consumed by the project's working group are only a fraction of all the costs it incurred to the whole office.

II.6 The organisation of the project

29.   The organisation of this project was different from the traditional project organisation of Statistics Finland. Usually there is a steering group, project manager and a working group. This structure was seen inadequate for the implementation project, so an adapted model of the traditional project organisation was introduced.
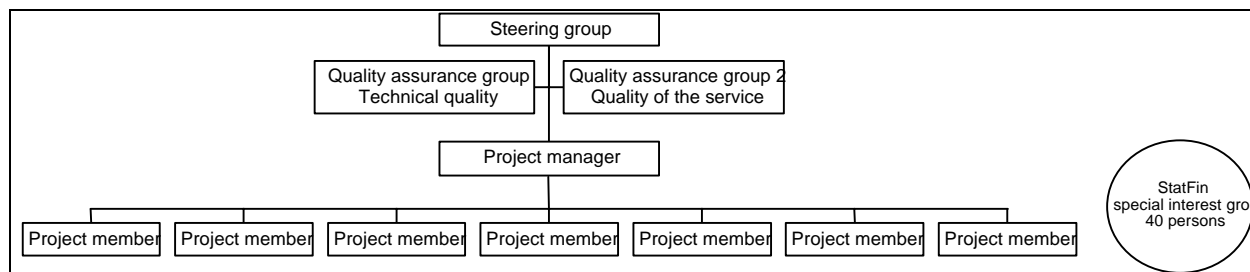

Figure 3. The organisation of the StatFin2000-implementation project.

30.   The project had a steering group consisting mostly of directors and very experienced experts. The project manager was chosen from the Information Technology

Services/EDP-methods unit. Quality assurance groups were introduced to report on the quality issues to the steering group and to the project manager. The members of the working group were experienced metadata- and IT-experts selected from IT-services and data administration department. The StatFin special interest group was a selection of 40 people from the statistical departments. They played a key role in exchanging information, distributing knowledge and promoting the system within the office - without their contribution the project would most likely have failed from the very beginning.

II.7 Schedule

31.   The Project began officially on 1 January 1999. However, some preparatory work was done before that, which enabled a smooth and rapid start for the project. The deadline for completion of the project was set to be the 31 December 1999.

32.   When the project was set, there was a set of milestones that formed the skeleton on which the projects working packages were built. The milestones were the following:

* Internal service must be up and running with usable data by January 1999;
* Public free-of-charge service must be up and running, with a significant amount of data by 15 May 1999;
* Public chargeable service must be up and running by the beginning of the year 2000.

II.8 Work packages

33.   As the project was to introduce an office-wide infrastructure change, there were three main work areas:

* Creating the technical infrastructure;
* Servicing and designing the service-concepts;
* Communication.

34.   These areas were divided into work packages. The work in most of these packages was concurrent and sometimes even overlapping. The work packages per work area were:

- Creating the technical infrastructure

* Purchasing and installing server-machines (2);
* Localising (definition of concepts and translation) software (StatLineSuite) and manuals;
* Developing conversion (import, export) programs to integrate StatLine with existing tools and infrastructure;
* Developing a web-site with Statistics Finland's look-and-feel;
* Finding a method to enable charging for the use of the data;
* Developing and automating the process of updating the data;
* Developing tools for usage-reporting and log-analysis.

- Servicing and designing the service concepts

* Designing the structure of the service;
* Integrating the service with the web-pages of Statistics Finland;
* Developing feed-back utilities;
* Following up on the use of the service and reporting on it;
* Developing best practices for service administration and administrating the service;
* Compiling a quality handbook on the service and immediately related processes.

- Communication

- In-office communication, presentations;
- Training, consulting, advising;
- Establishing a discussion forum and to actively use it;
- Customer relations, reaction to feedback;
- Marketing the service, participating seminars, promotion events, presenting papers;
- Cooperation with CBS, participating the work of the SOS consortium;
- Communicating the best practices and process descriptions.

II.9 The outcome

35.   The project succeeded in almost every task it was given. Statistics Finland now has an internal reference data warehouse, public statistical www-service (the StatFin statistics service) and technical infrastructure to build chargeable services using the StatLine technology. We also have over 100 persons trained to use the system in order to produce material (data and metadata) to the databases.

36.   The databases themselves are quite large and growing fast. The public service contains approximately 100 tables totalling 30 million cells (15.11.1999). The flow of feedback from the clients is constant, positive and growing. All feedback is stored in electronic form for processing and analysis.

37.   Other official statistics producers in Finland have shown a great deal of interest in the system and negotiations are under way to encourage them to participate in filling the services with data from all areas of society. This is a very positive thing because one of the roles of Statistics Finland is to coordinate the production of statistics in Finland and, with an attractive output tool, shared by many producers of statistics, this coordination and integration is clearly visible and useful to our clients.

II.10 Essential experiences

38.   The support (both financial and principal) of the high-level management is vital to an office-wide project. Without this kind of commitment, an office-wide change is impossible within one year.

39.   In-office communications are essential to overcome the resistance to change. People should always be aware of the impacts of the change on their workload. It is also important that everyone involved feels that the change brings significant advantages both for themselves and for the whole office.

40.   The StatFin special interest group was very useful in communicating between the project and the statistical departments. It was a way to establish two-way communications within the office thus enabling fast reactions to feedback and suggestions.

41.   Training a critical mass of in-office users (the content providers for the database) makes it easier to fill a warehouse with data. Training alone is, however, not enough: there must be a constant person-to-person support available in order to make the process of filling a database as streamlined and secure as possible.

42.   Distributed coordination is not always a good thing - at least some kind of an editorial board would be needed to coordinate the contents and structure of the service. Totally centralised coordination is still not the answer as it lowers the degree of commitment and makes the task of filling and updating the database look like a task for this central coordination body.

43.   When introducing a new system for dissemination, it is essential to integrate it with existing tools within the office. This reduces the workload when filling the system with data thus making it easier for the system to be accepted as a common tool.

44.   Rules and regulations are needed to lower the threshold for accepting the system, but in the long run the system must also prove to be an attractive tool, otherwise it will be rejected.

45.   Customers see WWW-based systems as more easy-to-use than the mainframe-based ones. This, however, requires the WWW-based user-interface to follow the mainstream design rules of similar systems.

46.   A huge collection of data disseminated free-of-charge is not a threat to a statistical office, but a significant way to build a positive image: content provider in the information society.

47.   On-line dissemination for a target group such as the Finnish people requires very thorough explanations and descriptions of the data in order to avoid misunderstandings and to make the data suitable for professional use. The explanations and descriptions should come from centralised metadata systems; otherwise, describing the data will be a huge burden for statistical departments.

II.11 The future

48.   The StatFin-service, as the other StatLine-based data warehouses in Statistics Finland, will continue their growth after the introduction and implementation project has ended.

49.   The system is in use in everyday production, but there is still a need to develop further, enhance and standardise the interfaces and streams for data and metadata, in order to seamlessly integrate the StatLine system into our production processes. This and reacting to the constant flow of user feedback will be a significant challenge for the coming years.

References

M. Saijets, P. Toivonen, S.I. Björkqvist, M. Mäkinen, R.Syvänperä, K.Palteisto, J. Kuosmanen (1999): Data architecture at Statistics Finland, Information Technology Services, Statistics Finland. Paper available in English only.

W. J. Keller, J.G.Bethlehem (1998): Between Input and Output, Proceedings of the NTTS'98 seminar, Sorrento, Italy, 1998.