Organisation de Coopération et de Développement Economiques
Organisation for Economic Co-operation and Development

15-Mar-2004

English - Or. English

STATISTICS DIRECTORATE
STATISTICAL INFORMATION MANAGEMENT AND SUPPORT DIVISION

STD/SIMS(2004)6
Unclassified

One Source Serving Multiple Needs - Easing the Data Collection of Recurrent Users, Statistics Denmark

OECD EXPERT GROUP ON STATISTICAL DATA AND METADATA EXCHANGE

1-2 April 2004
Château de la Muette, Paris

English - Or. English

JT00159993

ONE SOURCE SERVING MULTIPLE NEEDS – EASING THE DATA COLLECTION OF RECURRENT USERS

By Annegrete Wulff, Chief Adviser Electronic dissemination
Statistics Denmark

Background

The *StatBank* is Statistics Denmark's public database on the web. It covers very detailed but aggregated statistics from all subject areas and has a geographical detail down to municipality level where relevant. It contains 900 big multi dimensional tables with a total of 2 billions figures (excl nil). Since January 2001 the StatBank has been free of charge.

Behind the StatBank is the production database, the *Sumdatabase,* which is only accessible internally.

The Sumdatabase and the StatBank are both based on a common metadata model, a model developed and maintained by Statistics Sweden with input from Statistics Norway and Statistics Denmark. The internal interface from the SumDatabase is PC-AXIS-sql from Statistics Sweden, while the Internet interface from StatBank is in-house developed by Statistics Denmark - including components (dll's) from Statistics Sweden.

Dissemination Principles

Publishing is based on the following basic principles:

- the Sumdatabase is the foundation of our publications;
- there is one access point to electronic dissemination of statistics and paper publications from Statistics Denmark: www.dst.dk;
- dissemination should address well-defined target groups or types of usage;
- in the dissemination of statistical products, only one foreign language, English, is used by Statistics Denmark;
- StatBank Denmark is to be the place for all official statistics whether the data are collected and processed by Statistics Denmark or others;
- basic statistical data are available on the web site free of charge; additional statistics are generally published against payment;

- new figures from Statistics Denmark are published simultaneously at 9:30 a.m. in StatBank Denmark and News from Statistics Denmark (The News Release);

- certain statistics that are deemed to be of minor importance to the media or other large user groups are not published in News from Statistics Denmark, but only in Statistical News and Statbank Denmark.

Loading of Data and Metadata

The SumDatabase is an Oracle database. Data files are loaded using a simple flat file that can be created from several software tools. The file is created in the subject matter divisions on the basis of descriptions delivered by the databank staff. Loading of data into the Sumdatabase is a decentralized task carried out by the subject matter divisions.

Each load is checked for common errors and a log-file with load results will be e-mailed to the update person.

Governance of metadata is centralized in order to secure coherence across statistical areas and time. Metadata are split according to the nature of the information: common metadata used across the databank such as subject areas, statistical variables, value codes and texts etc. are created and updated centrally by the databank staff using a web tool called SumTimes. Metadata specific to a particular table, such as footnotes and contact person are created and updated by the subject matter divisions using another application, SumTools. Telephone number and e-mail address is added by matching against the employee administrative register. The subject matter division can only reuse existing metadata to create new matrices.

Privileges for editing and checking data and metadata are managed from a central database which restrains all of the above mentioned tools.

Metadata- and load software are in-house developed using Visual Basic, ASP, XML and Oracle procedures.

Release of Statistics

Statistics Denmark releases data at a fixed time once a day. There are no exceptions concerning access to first release data. No privileged users can get the data earlier than others. The procedures below concern the first release of data. This is crucial and follows the principle of concurrent access to official statistics for all users: The Parliament, the press, the Ministries, the Central Bank, the student and the international organisations get access at the same time. This is also the case with the News Release on paper and electronic publications. This is controlled in the same process because publications are produced directly from the Sumdatabase, too.

- Data are released at 9:30 AM regardless of medium. The electronic version will in most cases be more detailed than what is released on paper;

- All the electronic versions of the first releases are controlled by automatic procedures.

When the subject matter divisions load data in StatBank, they have to enter a date of release at the same time. Usually they will load data 1-7 days in advance of the release. Every night a job is run centrally checking if the date of the loaded data equals the actual date in the calendar. If so, the data are replicated to an Internet database server behind the firewall. The replication starts at 2 am and is ready by 5 am. At

exactly 09:30 am there will be a switch between the active database server and the one that was loaded during the night. Their position changes and the former public database will now be the internal one for replication the following night. Because of the switch mechanism which is regulated by a satellite controlled clock, even very big tables will be ready for release exactly on time.

Export of Data

The StatBank is the public database in Statistics Denmark. It serves a variety of users with a variety of needs. Someone wants to have a small table displayed on the screen, others need to make further calculations and still others import the files directly in applications for automatic update of a databases. A logging of the preferred output format shows that more than half of all 1.2 millions retrievals in 2003 ended on the computer screen as an HTML table without any further download. Export formats from StatBank counts over ten different types but Excel is the clear favourite with 80 % of all downloads. PC-AXIS is second. The user can in this case conveniently manipulate data before probably passing it on to another tool in his own environment.

- HTML table on screen (65 % of all *retrievals*)
- Maps on screen (6 % of all *retrievals*)
- Excel (80 % of all *downloads*)
- PC-AXIS (13 % of all *downloads*)
- DBase (3 % of all *downloads*)
- CSV-files (2 % of all *downloads*)
- SAS (1 % of all *downloads*)
- Time series format
- GESMES/TS
- XML

The GESMES/TS format is by Eurostat, NSIs and several International organisations agreed to be *the* standard format for exchange of data. It is not expected to be of interest to the normal end-user. That is why we do not give external users coming via Internet the possibility to download in GESMES. Statisticians within Statistics Denmark can choose GESMES/TS as an alternative.

XML is also for internal use at the moment, but it is envisaged to be highly important to professional users and redistributors in the future.

While the user is still online to the StatBank he can get a lot of descriptive metadata along with the retrieved table. Most important is a link to the quality declaration that describes every statistics. It contains information about the source, collection method, reliability, typical use of the statistics, etc. Other metadata are links to pdf publications analysing the specific release and finally footnotes connected to the retrieved table. Depending on the chosen download format the user gets more or less of this information. The links will not be downloaded, but in PC-AXIS and Excel the footnotes are included. This is possible because the footnotes are stored in the databank. However, links – for instance to a classification- can be included in a footnote.

Receipt of Data, Alert Services

Access to statistics goes via the StatBank user interface on the web.

By the end of 2003 more than 45,000 users have registered. Moreover, an unknown number of not-registered users are active, as the registration is not mandatory. However, the ones who register will have some advantages:

- they can retrieve larger tables – max 50,000 cells instead of 1,000;
- they can personalize the presentation of StatBank in accordance with their preferences (default formats, display the most used tables or the latest updated, automatic log-on);
- they can save queries and create selection lists for re-use;
- they can subscribe to *data shooting* and have their saved tables sent in an updated version at 9:30 the day of release. Free to choose an export format;
- they can have a news letter mailed.

Data shooting is a service that the registered users can subscribe to. There is a fee of 80 Euros a year which gives the right to an unlimited number of data shoots.

The user defines the query and the export format. As a new feature we will this year introduce Excel web query as an alternative within the data shoot concept. It means that once the user has got his data in Excel, he does not need to visit the StatBank to get updates. Data shoots from several StatBank tables can be combined in one Excel sheet by the user and additional calculations and manipulations can be made. In Excel he indicates whether the Excel sheet shall be updated “on request” or every time it is opened. In such a case, he is guaranteed that it is always the latest data from the StatBank he is working on. This can of course be combined with a data shoot notice at 9:30 as well.

Some users have already developed this functionality on there own, OECD being one example. We have no knowledge of their selections and we can therefore not protect them from changes in the databank that may lead to mishmash in their retrieval. We can on the other hand take care that subscribers of data shoots will be informed and we can make proper interventions.

It is our belief that data shoot in different formats will be very useful to professional users and international organisations.

Actually it is the same mechanism that will be used to automate the exchange of data in GESMES/TS with Eurostat. The responsible subject matter person in Statistics Denmark has to retrieve his data as a saved query (the first time). Doing so he selects GESMES/TS as the preferred format. A form which has to be filled in with the needed metadata information according to Eurostat requirements will then pop up.

From then on the normal procedures are followed: a program runs every night checking if a saved query in GESMES/TS matches an updated table in the StatBank. If this is the case the file is retrieved, saved in GESMES and can be sent to Eurostat (or other international organisations) 9:30 am in the morning.

For the time being we still have a few manual procedures in the process, but it is expected to run automatic in mid 2004.

Reporting to International Organisations

Statistics Denmark is responsible to report to several international organisations: UN, OECD, Eurostat, ECB, IMF,…. We receive questionnaires in many different formats and the requested data are often only slightly different from one organisation to the other. Contacts regarding international reporting are often between counterparts of the same profession and it has therefore been difficult to get a clear overview of the total. Statistics Denmark aims at centralizing all data in one place: the Sumdatabase. It is stored at an aggregated but very detailed level that makes it possible to serve a lot of different needs. Not everything in the Sumdatabase will be made public in the StatBank. Some data may be of interest only for a few users (this might be an international organisation) but as long as data are in the Sumdatabase we can use the functionality developed for automatic sending of data – or we can give restricted access to data in separate parts of the StatBank to authorised users.

These are functions not yet available. We hope to further exploit them in the framework of the cooperative work of SDMX, sharing of data and building of a common glossary. It is our expectation that international organisations will manage to retrieve data themselves from our database in a proper way, eventually via data shoot in some form. The OECD manages to do this already today.

GESMES/TS is the standard format for data to ECB and Eurostat. Therefore, Statistics Denmark has included this as an output from the database. For efficiency reasons, we do not want every single subject matter division to have to make their own solution for data exchange. We prefer a centralized solution where one tool is used by everybody to produce the output. In that case the advantage of recipient international organisations to agree on one single format would be highly appreciated.

Serving the Government

As well as international organisations are important users, the central and local administration and the Government are certainly so too.

The administrations have retrieved data from our databank since the very start of the online databank era back in 1986. Still today they are among the very frequent and satisfied users.

The Government has some particular needs which cannot be managed with publicly available statistics. They have a need to analyse consequences of new laws: changing the tax rules in a specific way will influence families differently depending on their income, seize, expenditures, subsidies etc.

For that purpose Statistics Denmark has established a "*Law model*". This is a tool, supplied by Statistics Denmark, only for use by ministries. The database used for the calculations contains anonymised micro data (on individual persons and households) with a large number of characteristics corresponding to parameters of the legislation. In order to minimize risk of identification, the "model populations" cover only a small random sample of the total population.

Serving Researchers

Researchers constitute 24% of the StatBank users. However, from the StatBank they cannot access micro data often needed in order to analyse the society. Confidentiality principles have to be followed and these data are not available from the public database.

That is why Statistics Denmark offers another solution. A research or analysis environment can apply for an authorisation from Statistics Denmark.

Access is not granted for all datasets; particularly sensitive data (e.g., data on crime) are excluded from the scheme and data on enterprises are assessed carefully to avoid any problems of confidentiality. It is emphasised that the data consist of samples. If the researchers request access to total populations, the content of variables must be limited. The individual cases are assessed by a steering group consisting of the Directors of Statistics Denmark. The scheme will be assessed regularly by the Board of Governors of Statistics Denmark. The technical solution is based on the use of the Internet. Like before, the data remain on the servers of Statistics Denmark.

Communications via the Internet is encrypted by means of a so-called RSA SecurID card, a component that secures Internet communications against unauthorised access. In practice the researcher rents a password key (a token) from Statistics Denmark. The token ensures that only the authorised person obtains access to the computer system.

The functionality at external access is largely the same as on the PCs under the on site arrangement. Printing and data transfer options are not available. Printouts are sent to the researchers by e-mail, logged at Statistics Denmark and checked by random sampling by Research and Methods.

The external electronic access by researchers is a valuable step forward for everybody involved. Statistics Denmark remains in control of the data, which are not handed out, and the researchers can work with the majority of the datasets from their own workstations.