Distr.
GENERAL

CES/AC.71/2004/6
12 March 2004

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE (ECE) CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS)
(Geneva, 17-19 May 2004)

Topic (i): Web technology in statistical information systems

WEB-BASED DATA DISSEMINATION SERVICES OF THE STATISTICAL INFORMATION SYSTEM GENESIS

Contributed Paper

Submitted by the Federal Statistical Office, Germany[1]

I. INTRODUCTION

1. Internet as a medium to transmit data between administrations, business and citizens and the use of web-technologies and standards to support the corresponding working processes have rapidly changed the process of communication between administration, business and citizens. In Germany the Federation will be offering 376 services over the Internet in the framework of an e-government initiative called BundOnline 2005 until the year 2005.

2. A long tradition and experience in using IT to support working processes such as data collection or data dissemination exists in statistical offices. It is, therefore, not surprising that these offices began to use and to draw benefit from web technologies very early.

3. In the Federal Office for six centralized statistics, the respondents are asked to answer the questionnaire via the Internet or to upload a file with their figures (e.g. external trade, distributive trade, cost structure). Since the beginning of 2004 we have been implementing one common application for the offices of the Länder and the Federation called IDEV, which shall handle the data collection for centralized and decentralized surveys.

4. In the area of cash statistics, web-technology was used to implement an pilot application, which covers all steps of the working process to build up that statistic, e.g. survey management, data collection, data editing, data aggregation and handling of metadata. Through the use of web technology the subject matter statisticians in the decentralized offices of the Länder are able to use the application, which has to be maintained in one office only.

[1] Prepared by Ernst Schrey, ernst.schrey@destatis.de

5. In the field of data dissemination and publication there are three different applications, which are based on web technology. First, is the homepage of the Federal Office, which contains a lot of static information and serves as entry to other services using the technique of HTML. Second, there is the statistic shop, where printed or electronic publications can be ordered by the user via the Internet. The goods can be downloaded or disseminated by traditional post. For some products the user has to pay a charge; paying by electronic cash is possible. This service was set up using a commercial shop software. Thirdly, there is GENESIS, which covers different services to access to data and metadata of the statistical information system GENESIS online or via a delivery service. The information transmitted is dynamic, which means it is tailored directly to the user's needs.

6. The last point will be discussed in more detail later in the paper.

II. DISSEMINATION SERVICES BASED ON GENESIS

A. Data model

7. Some aspects about GENESIS have been reported already at the MSIS meeting in 2001. The paper therefore will focus on the data model as it is important for the understanding of the dissemination services.

8. GENESIS covers the database with statistical figures and a system of metadata including information on surveys, variables and their items, rules for deriving variables, units of measurement and the existing frames for standard tables. The thesaurus includes all relevant keywords required to access the information designed for the general public. Information on the evaluation system such as table definitions, result tables and results of retrievals are also stored in GENESIS.

9. The logical data model is designed as a data cube, where each dimension (axis) is represented by a statistical variable (fig. 1 shows a cube with 3 dimensions); each point on an axis represents an item of the appropriate variable.

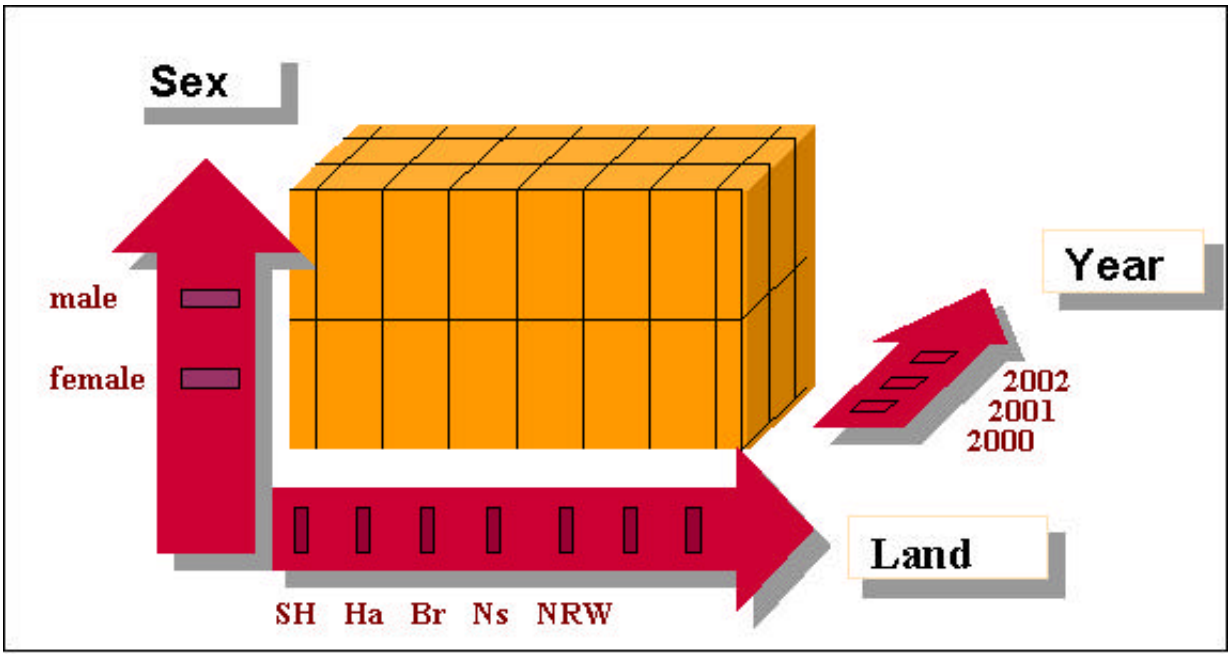

Figure 1: Data cube built by dimensions Year, Land, Sex

10. Every cell in a cube is defined by the combination of the corresponding items and contains one or more value variables like local units, persons employed, wages per month, etc.

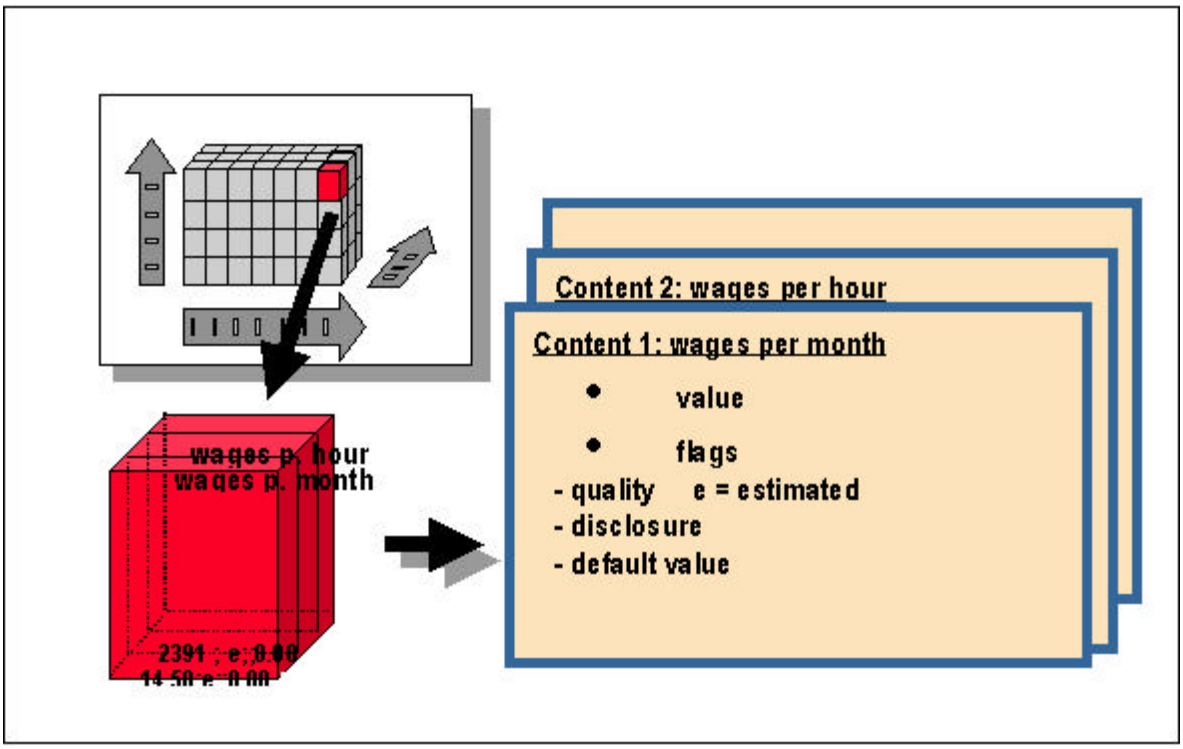

Figure 2: Data cell built by value variables "wages per month", "wages per hour" with additional information

11. Based on these data cubes the GENESIS table preparation component works to either prepare tables fully in line with user requests or use ready-made, i.e. standard tables offered by the data base provider. A table assistant is available o the user to obtain a tabulation proposal for a selected data cube.

B. Use scenarios for the use via internet

12. The experience gained in the last years in disseminating data on the basis of the former information system STATIS-BUND, the feedback we got from several discussions with users and two larger user conferences led us to design the following use scenarios:

- anonymous user with little knowledge in statistics, just wanting to get one figure to look at; e.g. number of female persons in Hessen;
- student, who needs data from different statistics to do some calculations and analysis for a paper, i.e. there is a need to transfer the data as a file in a machine readable form;
- periodical (monthly, quarterly) access to the same data, without the necessity to retrieve and select data every time from the scratch; e.g. data are needed to maintain the user's (research institutes, companies, information broker) own database;
- data are needed at the date of the publication, in some cases before (Federal Bank, Ministry of Economics).

13. Besides that there are some general requirements derived from the fact of world wide access

- Description of data (metadata) are very important
- Multilingual user interface
- System should be available 7 days a week and 24 hours a day

14. As a consequence of the general dissemination policy of the Federal Office, some dissemination services shouldn't be free of charge.

C. Dissemination services

15. Based on the use scenarios and the general requirements we differ between five types of users / services (guest, customer, premium user, data delivery service, privileged user).

16. To simplify using the system, the “guest” has only one kind of view on the data. The complex structure of the data cubes is hidden behind a data object called table. These tables are predefined by the dissemination unit of our office and stored as frames in the database. Retrieving information by a tree structure of statistical subjects or by catchwords, the user is led to a number of tables, which contain figures with regard to the retrieval parameters.

17. For each table the header and the columns and rows of the table - described by variables and their items - are shown to the user. If the proposal meets the user’s needs, it needs only a click on the button “preparation” to show the figures to him. Otherwise he can change the table by suppressing some items, e.g. select a time period or a part of a classification, but it is not possible to delete a complete variable. The results can be downloaded directly in different formats too.

18. Additional to the facilities of a guest the “customer” is able to store the results of his retrieval and selection as an own named table (frame). These tables can be used for data access and download without repeating the retrieval and selection process. Secondly he is allowed to access tables, which are not free of charge.

19. The type of “premium user” was created for research institutes, companies, etc. It includes up to ten user licenses as customer and the data access is not restricted to the predefined tables. With an additional export function the premium user is able to select and download data directly out of a data cube. He therefor needs too the ability to access the metadata describing the data cube in general and its structure (variables, items). The amount of data that can be downloaded in that way is limited per session by reason of performance.

20. If a user needs a larger amount of data once or periodical he may give an order to our office to deliver a fixed selection of data cubes to him. The delivery is carried out at different dates, depending on the date of update of the cubes. The data packages are prepared by staff of the Federal Office and stored in a file system on the so called “delivery server”. Then the user will be informed by e-mail and can download his data from there. As described later the whole process is automated and needs only little handling by staff. If a package is too large to download, it can also be delivered by CD-ROM.

21. Direct access to the basic information system GENESIS from outside is allowed to some “privileged users” (Federal Bank, Ministry of Economics) only. They can use the complete toolbar of the table preparation components and all data cubes like internal users. In contrary to the above-mentioned types of users who access via the internet, privileged users are restricted to use administrative networks.

D. Data objects and formats

22. Corresponding to the different needs of the users they have access to two different data objects: table and data cube. Depending on the data object which is accessed the user gets different types data, metadata and logical / technical formats.

23. For example, a company that wants to update its own database will choose the so-called cube-export format. Besides a lot of metadata describing the cube in general like survey, variables and so on the file contains one record for each cell of the cube with the subject-keys of the items of the variables and the content of the cell (value(s), flags).

24. Every part of the file (survey, variable, data cube, data cube axis etc.) consists of one record starting with the character "K" and one or more records starting with the character "D". The K-record defines the structure of the following D-records containing the data. From the technical point of view it's a CSV presentation.

25. Some records are listed below (Fig.3) as an example. They show parts of the cube-export for the data cube shown in figure 1.

```
K;ERH;FACH-SCHL;KTX;SPR-TMP;SIEHE-FACH-SCHL;LTX;ERL                    survey
D;62321;Verdiensterhebung i.prod.Gewerbe,Handel,Kreditgew.;N;;
"Verdiensterhebung im Produzierenden Gewerbe, Handel,
Kreditgewerbe und Versicherungsgewerbe";
"Für die Monate Januar, April, Juli und Oktober liefert die
.
.
.
K;DQ;FACH-SCHL;GHH-ART;GHM-WERTE-JN;SPR-TMP;REGIOSTAT;EU-VBD;      „data cube“
GENESIS-VBD;ERL;"mit Werten
D;62321LJ001;;N;N;N;N;J;

K;DQA;NAME;RHF-BSR;RHF-ACHSE;ERL                                       „dc-axis“
D;GES;2;2;
D;DLANDB;1;1;

K;DQZ;NAME;ZI-RHF-BSR;ZI-RHF-ACHSE;ERL                              “dc-time axis“
D;JAHR;3;3;

K;DQI;NAME;ME-NAME;DST;TYP;NKM-    STELLEN;MAX-SBR;ERL                 „dc-cell“
D;VST007;EUR;FEST;DURCH;2;0;
D;VST008;EUR;GANZ;DURCH;0;0;

K;QEI;FACH-SCHL;FACH-SCHL;ZI-WERT;WERT;QUALITAET;GESPERRT;WERT-VERFAELSCHT
                                                                  „cell content“
D;GESM;01;2000;14.50;e;;0.00;2391;e;;0
D;GESM;02;2000;16.25;e;;0.00;2687;e;;0
D;GESM;03;2000;15. 98;e;;0.00;2524;e;;0
D;GESM;04;2000;16.26;e;;0.00;2631;e;;0
```

Figure 3: Example cube-export format

26. While this generic format needs some programming to be handled further the .XLS format can be used to work with tables without. The following summary gives an overview on objects, services, metadata and formats.

object	table (data derived from cube data by summing or selection)	cube export (cube data for selected cells and cell contents)
service	guest / customer / premium user	premium user, data delivery service
metadata	subject key and/or texts (short version)	subject key and/or texts, description of survey,...
format	.XLS, .CSV	cube-export as CSV

Figure 4: Data objects and corresponding formats

E. IT architecture

27. The services are based on a client/server architecture. The software is built in 3 tier architecture using DBMS ADABAS, programming language Natural, Entire X as middleware and Java, XSSL for the frontends. The software of GENESIS is installed twice:

- on the basic server system (Siemens BS2000 OSD 3.0) for the data and metadata import, internal use of GENESIS by subject matter statisticians, privileged users, data delivery service (data export) and as source system for GENESIS-Online
- GENESIS-Online (UNIX Sun Solaris) for external access via the Internet (guest, customer, premium user) and data delivery service (user and order administration, handling of data packages).

28. We have chosen that configuration to get a higher level of security and scalability. It raises problems and needs capacity to maintain the second server and to keep the contents of data and metadata synchronized however. The latter is managed by exporting the data from the basic server system and importing it in the online system using facilities of GENESIS. Therefor the content of GENESIS-Online always is a real part of the basic system. The external user interacts with GENESIS via an web-server , using a XML-interface of GENESIS.

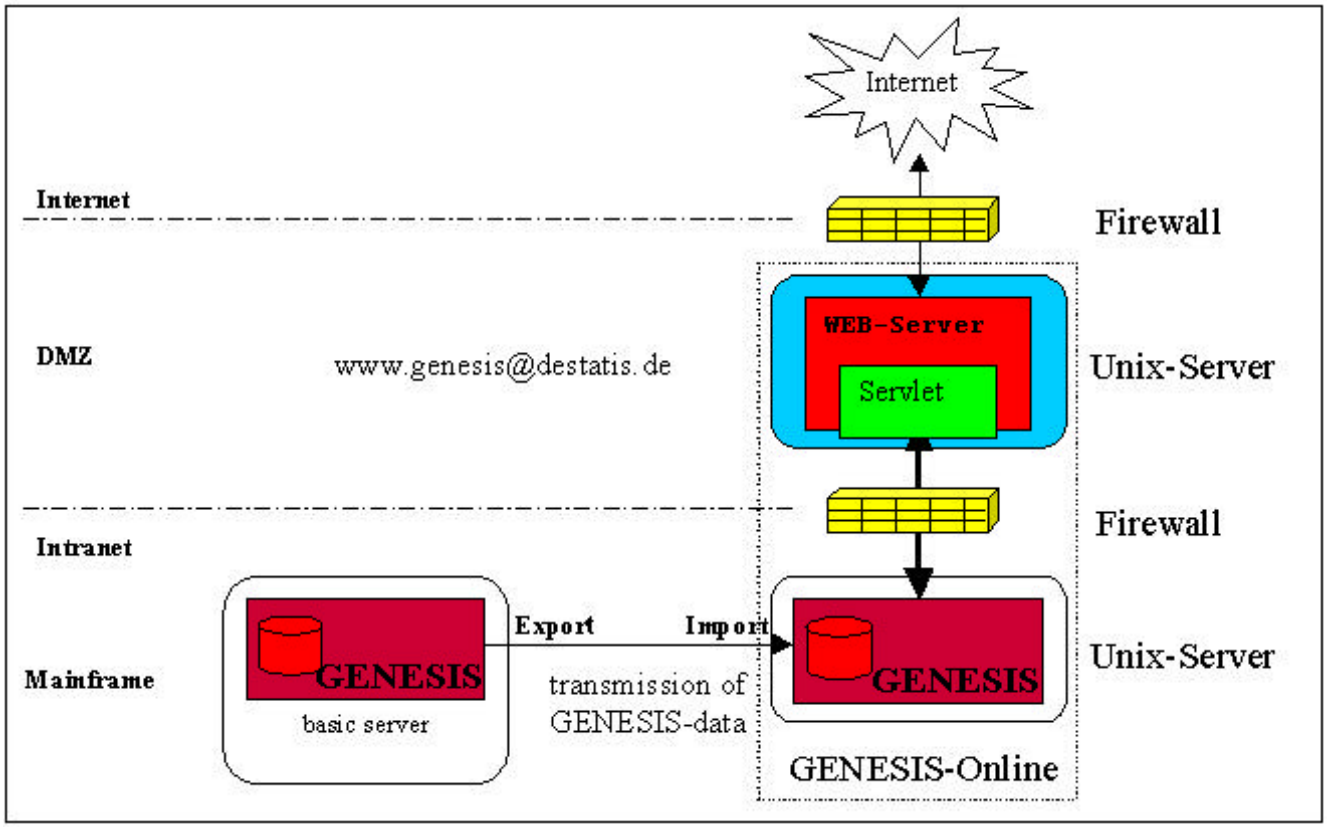

Figure 5: IT architecture GENESIS and GENESIS-Online

29. Figure 6 shows the interaction of different components and data flow for the data delivery service. To handle a new order the user support enters a corresponding record in the order file on the basis server system. When the data cube is updated, the order procedure generates a batch job to export data out of the cube and transfer the result to the online system or the CR-ROM writer (offline delivery). Within the online system an order-demon handles different export files of one order and user information to prepare a delivery package and transfer it to the delivery server. Checking user access and downloading is managed by the delivery server.

30. The administration of users is done in GENESIS-Online in common with other registered users (customer, privileged user).

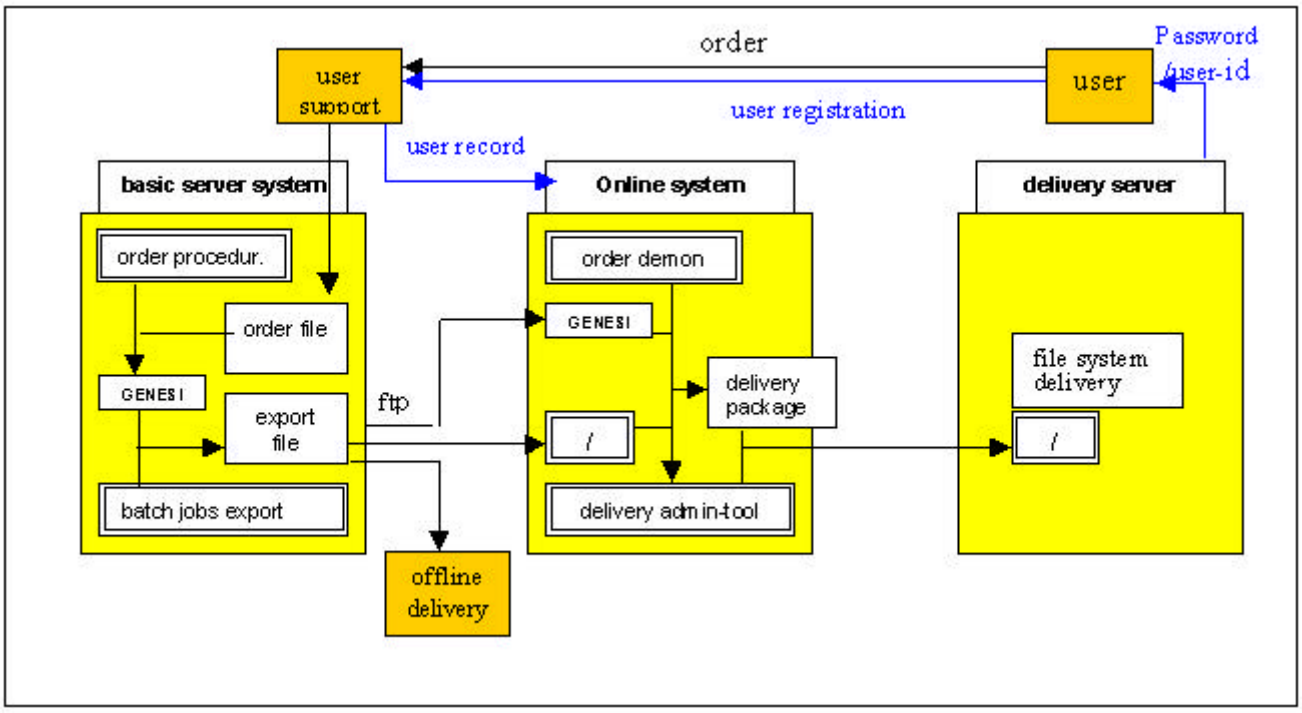

Figure 6: Data flow delivery service

III. CONCLUSIONS

31. Internet and web-technology provide new opportunities and chances to meet the demands of users of statistical data for the statistical offices. Nevertheless, it's a new challenge too to face the wide range of users, which reaches from the specialist in statistical matters to the layman in that field. Users are known or anonymous. The dissemination services should be available all the time and in different languages.

32. The different services described in the contribution were developed under consideration of our own experiences gained in the last few years with the former Information System STATIS-BUND and the many discussions and contacts we had with different users. They are formed as a toolbar of complementary services, as we are convinced that the variety of users needs a variety in services to build a user-friendly system in total.

33. Having finished this development at the end of last year we are aware already of new needs such as automating the access or delivery of data on the side of the user or to use the retrieval facilities of GENESIS Online in the framework of general portal to combine statistical data with geographical data to build up a geographical data infrastructure. For that purposes the use of web-services technology seems to be adequate, but a lot of questions concerning security matters, have to be dealt with.