

SEMINAIRE

E+; 3=!C

SEMINAR

STATISTICAL COMMISSION AND
ECONOMIC COMMISSION FOR EUROPECONFERENCE OF EUROPEAN
STATISTICIANSDistr.
GENERALCES/SEM.43/23
24 March 2000

ENGLISH ONLY

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

(Riga, Latvia, 29-31 May 2000)

Topic IV: Improving data dissemination strategies

REFLECTIONS ABOUT DATA DISSEMINATION BY STATISTICAL AGENCIES

Invited paper

Submitted by the Federal Statistical Office, Germany¹

I. INTRODUCTION

1. Besides data collection and production, data dissemination is a basic task of all statistical agencies. In times of cuts in the budgets of nearly all statistical offices and the competition with private services, data dissemination gains in importance, as it is the point where the statistician comes in contact with his user and financier. In addition, data dissemination obviously is a measure for the statistician's work. In discussions about statistical offices these effects lead to a call for more output orientation. Without any doubt, statisticians have to attach more importance to users' needs, but one must keep in mind that there could not be a good data dissemination without proper data collection and production and vice versa.

2. This paper consists of two parts; the first part attempts to provide an abstract model of data dissemination; and the second part shows some experiences in that field in the Federal Statistical Office of Germany.

II. DATA DISSEMINATION AS A PROCESS

3. In the following, data dissemination as a production process, by which statistical agencies produce **data** and **disseminate** them to **users** will be reviewed. The goods produced by the statistical agency in that sense are not only the data itself, but also the way in which the data are distributed to the user (services). It is obvious that there are many different types of users (journalists, enterprises, ministries, scientists) and the kinds of data they want (tables, charts, microdata, even statistical expertise might be seen as "data"), as well as the media to be used for the dissemination (paper, telephone, diskettes, CD-ROM or

1 Prepared by Ernst Schrey.

online transmission). In addition, users have very different characteristics concerning technical know-how and equipment, financial power, knowledge about statistics and the time frame they want to receive the data.

4. Based on these elements, the data dissemination strategy and the decision about the production process necessitates a decision on a **bundle of products**, the **media** and **techniques** and tools to produce and disseminate it and a consistent **system of formation of prices**. This has to be done under the condition that the benefit function is maximised when considering criteria such as the characteristics and needs of the users, the needs of the National Statistical Institutes (NSIs) e.g. low cost of production and legal regulations. The benefit function should be fixed before, that means that a "political" decision has to be taken about the weights of the different criteria mentioned above. It is obvious, therefore, that some criteria are in contradiction to others (for example, the interest of the user in low prices and the wish to raise the receipts of the NSI).

III. EXPERIENCES IN THE FEDERAL STATISTICAL OFFICE

5. In the past, some of the above-mentioned criteria and conditions changed. The technical equipment of the users has also changed in computing power and the software available. The telecommunication connection between the user and statistical offices became very easy and is nearly "standardised" by the Internet. The NSIs are obliged to lower the cost of production and to raise their receipts. The requirements from scientists to acquire access to microdata are increasing compared with the past.

6. As an illustration of the above reflections about the data dissemination strategy, six examples of "products" will be discussed which are reactions to these changes; the considerations which led to their development will be described and some hints given to particular properties. They are:

- Timeseries-service (access to mainframe-based timeseries-databases via Internet),
- Konjunktur Aktuell (replacement of a printed publication about analysis of current business cycles by a Internet-based one),
- Statistik Shop on Internet (pilot-project on electronic commerce),
- Distribution of press releases by a satellite broadcast-system,
- Factual anonymous data - delivery of microdata to scientists,
- "Access" to microdata - co-operation with scientists by an exchange of the software-code.

These examples have been selected from different fields to show the broad range of data dissemination activities and what challenges the statistician has to face.

7. For each product, a description of the procedure and the application will be given. As dissemination strategy is in the focus of that contribution, examples for the different products are not described in great detail, but an indication will be given of where to obtain more information.

III.1 Time series service

8. The Statistical Information System of the Federation (STATIS-BUND) forms the basis for electronically storing, processing and disseminating data from a large variety of areas of official statistics. Currently, the system comprises about 2.0 million time series with monthly, quarterly, semi-annual or annual periodicity. To the extent permitted by the subject background, the data are comparable over longer periods of time. The longest series began in 1960.

9. Apart from the actual data, the system encompasses detailed documentation on related surveys, classifications, legal bases, conversion and estimation procedures, data quality and also other information relevant to interpreting the figures provided. The system functions as a dialogue application, its programme is run under the BS2000 operating system on a Siemens mainframe. The options of the dialogue application (using terminal or personal computer) were discussed in detail in the document entitled "*The Statistical Information System of the Federation as a working instrument for statistical experts and users of official statistics*" (CES/WP.9/284) presented to the February 1991 meeting of the Working Party on Electronic Data Processing.

10. The mainframe provides users with all functions required to analyse and assess the data of the central database. However, there are mainly two reasons in favour of providing external users not only with on-line but also with other forms of data access. On the one hand, dialogue access to the system has to be strictly monitored since the programme is run on the mainframe for the production of statistics of the Federal Statistical Office. As far as the user is concerned, this implies certain restrictions and requires a greater technical effort than is, for instance, necessary for simple dialling access. On the other hand, the data processing capacities have considerably increased among external users of statistical data. Therefore, the dispatch of user selected time series on disk has been a successful way of data dissemination for several years.

11. Besides, larger amounts of data, i.e. whole data packages, have been supplied on magnetic tapes and recently, also on CD-ROM. Even though the informative value of those means of data supply is equal to that of on-line access (selection can be made down to the level of individual series or periods), one aspect needs improvement: the procedure of ordering, compiling and mailing a disk still takes a few days. This period can be shortened by the new means of communications technology.

12. The first version of the time-series service was used from 1993 and its communications components were developed internally. Since 1996, the time series service is available via Internet and provides the opportunity to view all metadata of the Statistical Information System, to order an individual selection of time series data with the related documentation and to transfer the respective data on the user's computer. The metadata, which are updated at certain intervals, are directly stored on the Internet server and, in an automated manner, the time series requested are directly copied from the stock of original data of the mainframe and made available to the user.

13. Users have unlimited and free access to the metadata offered while time series can be retrieved by registered users against payment only. Registration is made once via the Internet. Users just have to click on the button *Register* and enter their personal data in the form. The data to be entered are the name and address including postal code of the user. Information concerning the institution, telephone, fax and e-mail is optional. In addition, the user has to print out the related unilateral use agreement, in which his personal data have been entered automatically, and send two signed copies of the agreement to the Federal Statistical Office. (For the time being, an electronic authentication procedure has not been introduced since settling all related issues would have considerably delayed the establishment of the whole service.)

14. Upon arrival of the above copies, the user will be provided with a user-ID and a password as a precondition for time series retrieval. Supplying time series comprises two steps, namely requesting and collecting the data. The user will obtain the information required to order data by searching the metadata. The easiest way is to mark the respective metadata and to enter the necessary information about the time intervals requested.

15. The system will then create a data file (order-file) including all elements

identifying the series requested. An experienced user can produce the above data file himself or alter an existing file. To this end, the system offers an editor. The order-files of a user can be stored to facilitate orders recurring at periodic intervals. Upon execution of an order, the user will be provided with the data and metadata in a compressed form in a specific directory of the Internet server. As in the request files, the user will be informed about the data accessible to him in the form of a list. The list also includes information about the amount of data provided and the price charged for the supply of data.

16. The data transfer to the user's computer will be initiated by marking the data requested in the list. The data will remain on the server for a certain period of time (about three months) so that they can be retrieved several times (e.g. in the case of data overrun or problems in data transfer). To decompress the data on the computer, to convert them into a conventional PC format or just to present them, the user may apply the FORUM programme provided by the Federal Statistical Office. Data may also be requested in the ASCII format directly. Performing the above steps of ordering and providing the data on the server takes about 15 minutes.

17. At the end of 1999, we had about 2500 registered users. In 1999, (in brackets the figures of 1998) they submitted 13433 (9076) orders for 798443 (582214) time series with 38347057 (27921093) values. About 200 customers cancelled their contracts in 1998 and 1999. The customers per field are: enterprises (45.8%); private persons (12.3%); universities and schools (12.2%), and consultants (10.2%).

III.2 Statistik Shop

18. Since March 2000 we offer as a pilot project the so-called "Statistik-Shop". By this means, we sell printed and electronic publications on the basis of a commercial e-commerce software. Till now there are 35 printed and about 60 electronic products on offer. Electronic publications are available in the file-formats of Microsoft-EXCEL (Excel97) and Adobe (PDF).

19. After registration, which has to be done once via the Internet, the user is identified by his user-id and password. He then "walks" through the shop with a shopping bag to place his orders. The retrieval is done by a hierarchical tree structure. After the user has finished his walk, he can order the contents of his shopping bag. The electronic products can be downloaded directly after the order and the printed matters are disseminated by the German Post AG. The order should be transmitted to the post within a maximum time frame of 24 hours.

20. Whereas the bill for the printed matter is sent along with the publication, the recipient of electronic products receives a monthly bill. Use of credit cards is in preparation, but there are some administrative questions to be clarified.

III.3 Konjunktur aktuell

21. "Konjunktur aktuell" was a printed publication of about 120 pages. It appeared monthly and contained basic information about conjunctures such as number of local units, employees, hours worked etc. for different economic branches and some figures about wages and labour market for the last six months. In addition, it contained a more detailed analysis for about 70 relevant time series (e.g. production index, turnover, index of orders for different economic branches) covering the last 36 months. The results of the seasonal adjustment such as trend-cycle, seasonal adjusted series, noise etc. (with value and changes from the preceding month or year) were shown in a table and as a graph. Some methodological comments were included. In this way, the user received a summary about the relevant figures of the economy and did not have to do the analysis himself.

22. The main problem with the printed publication, was to be up-to-date. Although the publication was produced by access to the central database (STATIS-BUND), which

contains the figures directly after production, the publication could not be completed before the last figures were received, as the time of production of the different statistics is distributed throughout the month. This could be considered as a contradiction to the name of the publication "aktuell" - that is up-to-date.

23. To overcome that problem, the data and results of the seasonal adjustment are published now via Internet and the dissemination of the publication in paper format has been discontinued. Thus it is possible to update the publication for each series separately. Furthermore, the presentation of the graphs is designed in such a way that the user can decide which components of the time series-analysis he wants to see in combination. In the paper version the number of components was fixed, as a diagramme of more than two components would not have been clear.

24. In spite of these advantages it should be kept in mind that moving to the Internet, which means data presentation on computer screen, has resulted in some other changes. So the cross-section comparison was omitted, the period for which data is presented is reduced to 24 months and the tables with up to eight columns for each series are split into two. Historical data that are available to the owner of the printed version by storing the publication is not available to the Internet user any longer.

III.4 Satellite system

25. Most of these services are so-called "pull" services, which means that the recipient of the data has to activate the dissemination process. There are other kinds of dissemination such as the dissemination of news items to news agencies. This is done normally by fax or mail, these are so-called "push" services. The problem was to ensure simultaneous release of highly market sensitive, economic data to agencies located in different cities. This could not be ensured when using the "ordinary" method of dissemination.

26. The Federal Statistical Office therefore made a contract with a professional satellite operator, who set up a broadcasting system. It uses the one-way VSAT (Very Small Aperture Terminals) technology to transmit the data via satellite. In this way, it is ensured that the subscribing agencies receive the data in the same second at all places in the country. To import the news and use the service, we only need a special workstation which is connected over ISDN lines to the computer centre of the provider. For further details, see the paper "The dissemination of news items via satellite" presented to the 25th Session of the Working Party on Electronic Data Processing (Geneva, February 1997).

III.5 Scientific use file

27. The purpose for this kind of dissemination is to satisfy the permanent growing demand on the evaluation and analysis of statistical microdata. The legal basis in Germany is a special regulation in the general law for statistics concerning microdata dissemination for scientific purposes (BStatG). Whereas it is forbidden to release individual data outside the statistical office in general, by paragraph 16 it is allowed to disseminate so-called "de facto anonymous" microdata to universities and institutes, which have the commission of independent scientific research and for a specific research project. "De facto anonymous" means that disclosure could be done only with a very high and uneconomical expense of time, money and human resources.

28. As a first step, the legal definition "de facto anonymous" given to the statisticians by law had to be made operational. A research-project was started to check against the risk of disclosure methods to make anonymous the microdata of the so-called microcensus (1% sample of the households in Germany). As the main result of the project, 4 methods can be mentioned:

- Sub-sample of 70%;

- Clustering of the regional attribute in every case;
- Clustering of those attributes, which show small figures in the univariate distribution (less than 5000 in the raised figures). In the case of microcensus these are year of marriage, profession, economic branches, hours of work per week.
- Some attributes are deleted totally (e.g. name and address) or transformed into an artificial one (e.g. number of households is replaced one to one by an artificial number).

29. Transformed microcensus data are called a de facto anonymous basic-file. Such a file is generated once for every survey. From it, for every research project a specific file - the so-called scientific use file - is compiled by selecting the requested attributes. The scientific use file is given to the scientist who has to sign a contract with the statistical office. Researchers who handle the data are liable for maintaining confidentiality and data have to be deleted at the end of the research project.

30. Scientific use files are a tool to support very special users with a very special product. The costs of preparing the product are very high and, until now, mainly borne by a ministry as part of a general research project. There is no uniform way to build anonymous data sets and basic files have been prepared up to now only for four population surveys. Nevertheless, most users are satisfied with this method, as the time to get the data and the price are reduced compared with the past due to preparing the basic file.

31. Similar work for economic statistics data is just starting. The procedure is described in "Wirtschaft und Statistik" ISSN 00436143, January 2000 "Pilot-project to simplify the use of de facto anonymous microdata".

III.6 "Access" to microdata

32. The demand of scientists to have access to microdata of official statistics meets the same needs as described above. As scientific use files do not exist in every case, the statistical office is testing another procedure, i.e. to perform the analysis for the scientist in the statistical office and to disseminate only the results. Similar procedures for the tabulation of microdata have been in use for a long time, but it was too expensive (the user had to pay for the programmers in the office) and was inflexible (every change in the table necessitated a change in the programmes). Cooperation has been mostly limited to tabulation, as more in-depth analysis was too complicated to harmonise.

33. Therefore a new form of cooperation was launched where the programmes are written by the scientist in an agreed standard software e.g. SAS or SPSS, tested with artificial microdata and transmitted to the statistical office. Then a member of the statistical office has to run the programmes using the original microdata and disseminate the results to the scientist. It has to be stressed, that this kind of "dissemination" or service requires a special IT-infrastructure and a trained statistician, who knows the software and who checks the results for problems of disclosure.

IV. CONCLUSIONS

34. The above examples show the diversity in data, the recipients of the data, their financial capacity and other conditions that have to be kept in mind when planning and deciding on a data dissemination strategy. As a result, there are services determined for one or more concrete users as in the case of scientific use files, others are aiming at a broad range of "anonymous" people or units like the time series service or "Statistik Shop". Whereas the user of the time series service is allowed to select a single time series out of a stock of about 2 million original data by a quasi-direct access to the mainframe, the user of the "Statistik Shop" receives an electronic

publication with tables and explanatory text, although both download statistical data from the systems of our office.

35. Nevertheless there are some common principles to highlight. First, a central database with up-to-date data and metadata should be the basis for different dissemination activities. In our case, the time series service, "Statistik Shop" and "Konjunktur Aktuell" are using the database STATIS-BUND by automated access. The second is the importance of Internet as a cheap, standardized communication system including its different services. However, it should be kept in mind that the possibility of disseminating data in electronic form to many recipients, who sometimes are not familiar with statistics, creates challenges concerning the documentation and metadata. It also increases the potential of comparing statistical data, which describe the same facts but are published in a different context. The third point is the format for the disseminated files. Until now we have been using HTML, EXCEL-spread-sheet, PDF and basic ASCII, but it will be interesting to follow up the progress of developments such as XML.