A STUDY OF THE SSD WEB LOGS FOR

SEPTEMBER 1999 AND SEPTEMBER 2000

By Svein and Joan Nordbotten
January 2001

Summary:

This study is based on registration data, provided by SCB, for users of SSD and log data for their use of the database in September 1999 and September 2000. The purpose of the investigation has been to gain knowledge about the users and their use of SSD to enable SCB to improve the statistical database and its services to the users.

The number of registered, external users of SSD increased from December 1999 to the end of September 2000, by a factor of 5. This increase was partly because of the cancellation of the annual user fee and partly because of the increased use of Internet by the public. Only a small fraction of the registered users, 7%, were active users in September 2000. Economic enterprises appear to be the heaviest users. While they made the most requests, "Other schools" is the category, which made most requests for output of statistics. During September 2000, Population, Commerce and Labor statistics, were most frequently requested.

Active users spent an average of 7 hours in SSD in September 2000, using an average of 3.1 sessions. A session has been defined as all the requests from a unique user during a day. The number of requests per session varied from 1 to more than 700. About half of the sessions included 10 or fewer requests. The time spent per session varied from 0 to 24 hours with an average 2.3 hours per session.

The current logging system has some deficiencies which make it impossible to investigate where users come from when visiting SSD, study their use of keywords in the searching process, distinguish individual users from concurrent users with the same user-id, and relate output requests to the specific tables in the database.

SCB should consider developing a logging system that permits reporting on the above aspects as well as a reporting system for management decisions that could compile central information about the statistical users and their needs at regular intervals.

Content:

1. Project contract 3
2. Proposal, discussions, and pilot study 3
3. Purpose of this investigation 3
4. Swedish Statistical Databases (SSD) 3
5. Concepts and terms 4
6. Users of SSD 4
   - *Who are the users of SSD?* 4
   - *Which categories do the users represent?* 6
   - *Where are the users located?* 7
   - *How active are the registered users?* 9
   - *Do the activities of registered users change over time?* 10
   - *How frequently do users visit SSD?* 11
   - *Which keywords are used?* 13
7. Requested and retrieved statistics 13
   - *Which statistics are requested?* 13
   - *What is the request frequency by user category?* 14
   - *Which statistics are requested by user category?* 14
   - *How are data retrieved?* 15
   - *What is the volume of downloaded statistics?* 16
8. Sessions in SSD 16
   - *How many sessions were recorded in September 2000?* 16
   - *What was the most frequent start request in the sessions?* 17
   - *How many requests do sessions include?* 18
   - *How does session length differ by user category?* 19
   - *What is the session row output?* 20
   - *From where are sessions initiated?* 21
9. Time spent in SSD 22
   - *How much time did the users spend in SSD in September 2000?* 22
   - *How much time is spent on sessions?* 23
   - *What is the time spent on requests?* 23
10. Conclusions and recommendations 24
11. Tables 27
    - *Table a: Registered users by months. January 1997 to September 2000.*
    - *Table b: Registered users by categories. September 2000 and September 1999.*
    - *Table c: Registered users by foreign country addresses. September 2000.*
    - *Table d: Registered users by categories and foreign/domestic addresses. September 2000.*
    - *Table e: Registered users by domestic regions. September 2000*
    - *Table f: Registered and active users by categories. September 2000.*
    - *Table g: Registered and active users by years of registration. September 2000.*
    - *Table h: Requests by user categories and statistical areas. September 2000.*
    - *Table i: Output requests by forms. September 1999 and September 2000.*
    - *Table j: Output requests by forms and categories. September 2000.*
    - *Table k: Sessions, active and registered users by categories. September 2000.*
    - *Table l: First requests by statistical areas. September 2000.*
    - *Table m: Sessions by size groups. September 2000.*
    - *Table n: Average session sizes by categories. September 2000.*
    - *Table o: Average row outputs by categories. September 2000.*
    - *Table p: Sessions by user addresses. September 2000.*
    - *Table q: Time spent by user categories. September 2000.*
    - *Table r: Sessions by time. September 2000*
12. Appendices 37
    - *Appendix A: Data received* 37
    - *Appendix B: Terminology* 39
    - *Appendix C: Log file for 16.april 2000 - An illustrative pilot study* 42

1. Project contract

This report is prepared according to contract 2000/2099 of November 20, 2000 between **SCB** and **SVEIN NORDBOTTEN** for a study of the user logs for the Swedish Statistical Databases (SSD).

2. Proposal, discussions, and pilot study

Svein Nordbotten proposed the study for SCB in April 2000. The proposal was discussed between SCB/Unit for Information and Publication (IP) and Svein Nordbotten (SN) in May. SN made a pilot study based on an extract for a single day (16 April 2000) from the SCB's general log file to clarify the aim and scope of a further investigation. A copy of the report from the pilot study is attached to the present report as ***Appendix D***.

A meeting was held in Stockholm September 4, 2000 between IP and SN. It was decided that the project should be carried out and that it should be limited to the logs for SSD with the aim to analyze the user population of SSD and its recorded behavior.

A brief second meeting between Lars Nordbâck, IP, and Svein Nordbotten, SN, was held November 9, 2000, for clarification of details. The contract for this study was dated 20 November 2000. It required that the project should be finished within year 2000 and a report submitted to SCB in January 2001.

3. Purpose of this investigation

The purpose of the project commissioned by SCB was to report on the:

- SSD user population at the end of September 2000,
- use of the SSD in September 1999 and September 2000,
- trends in the user population from January 1997 to September 30, 2000, and
- changes in use from September 1999 to September 2000.

4. Swedish Statistical Databases (SSD)

Statistics Sweden has published statistics on the net for a number of years. A pilot investigation indicated that the number of *different* visitors to the SCB web site a day in April 2000 was about *1.500* per day. These users made about *17.000* requests to the site, of which *2.000* were requests for entry to the database.

SSD was established on the Internet in January 1997. to make official statistics for Sweden more accessible, and easy for users to observe and download. In September 2000, about *800* different products or statistical tables were available in SSD. For the first 3 years, until the end of 1999, users had to pay an annual fee for access to SSD. From January 1, 2000, SCB has offered free access to SSD. For security reasons, all users still have to register before getting admission to SSD. This gives useful data for evaluation of the actual use of SSD.

Several options for output are available from SSD including simple statistics in lists and relational tables. The users get the option to get the requested statistic displayed at their monitors and/or downloaded as files in alternative forms for the users themselves to process with different tools.

The DB-log data used in this project were received on a CD-ROM from the SCB/IP. The files received are listed in ***Appendix A***. After concatenations, the files used were: the file of users registered from the beginning of January 1997 to the end of September 2000 (*Internetkund_personar.txt)*, and the log files for September 1999 and September 2000 (*Web_logfil_99_09.txt and Web_logfil_00_09.txt*). A 4th file (*Person_ internet.txt*) was used only for supplementing the user file with country addresses and for checking suspicious, individual records. Documentation of the log files was obtained from the IP division.

The files were preprocessed by several operations for the data analysis:

Cleaning
Only pure numeric *user-ids* and *user-ids* starting with an A or K preceding a number, were accepted. Other symbol strings were rejected because of uncertain interpretation. Some internal users from SCB were excluded as well as a number of obvious duplicates. This cleaning process resulted in *8.316* accepted registered users at the end of September 2000. From these users, *14.530* requests were recorded in September 1999 and 53.598 requests in September 2000. Since there may be users who were incorrectly excluded, the number in this report must be considered a minimum.

Adjusting
The two log files were adjusted to the user file, i.e. all requests in the log files from 1999 and 2000 were checked and only records with a user-id value in the user file were accepted.

Supplementing
The user file was supplemented with country addresses in an additional field from the file *Internetkund_person_ 00_09_01_09_30.txt*. The auxiliary file used for this operation did not cover all users in the user file. Users with no available address were assigned to location Unknown.

5. Concepts and terms

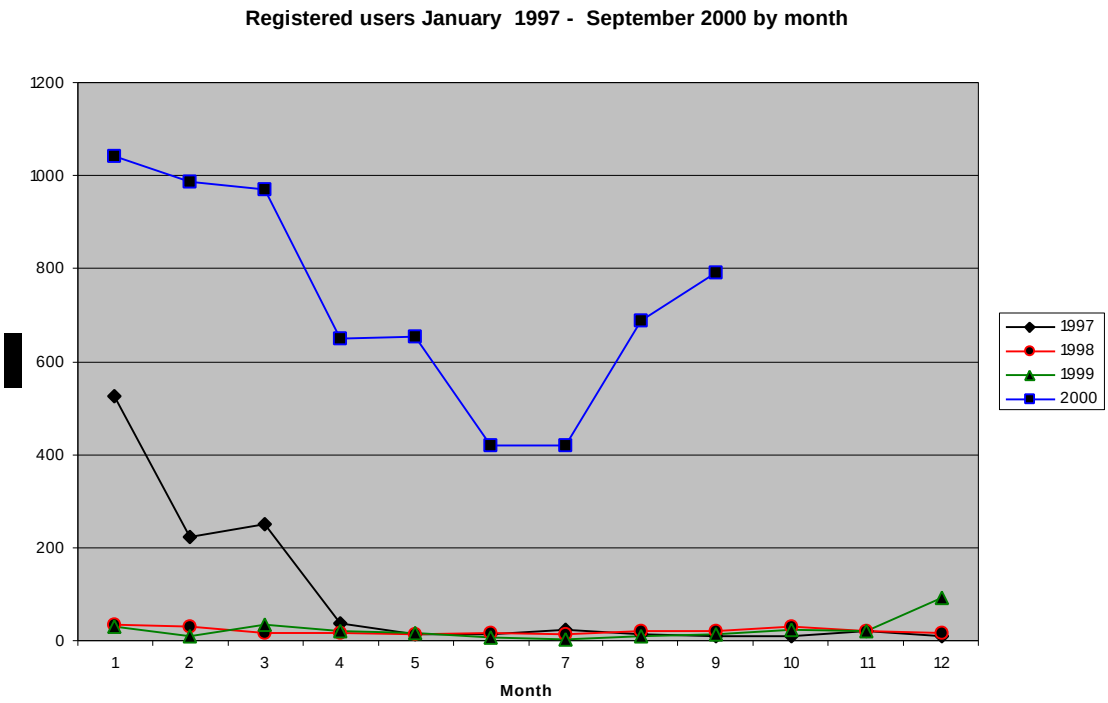
A number of special *concepts* and *terms* are introduced and used in this report. The most important concepts and terms are defined in ***Appendix B***.

6. Users of SSD

Who are the users of SSD?
For this project, the user population is limited to the registered external users at the end of September 2000. It was decided that internal users within SCB should not be included.

Users of SSD have been registered since 1 January 1997. In the 3 first years, the general rule was that the users had to pay an annual fee for access to the databases. From 1 January 2000, anyone who submits an application for a user identification and password gets free access. It is therefore of particular interest to study the development before and after 1 January 2000 to assess the effect of canceling the fee for access.

Chart 1:

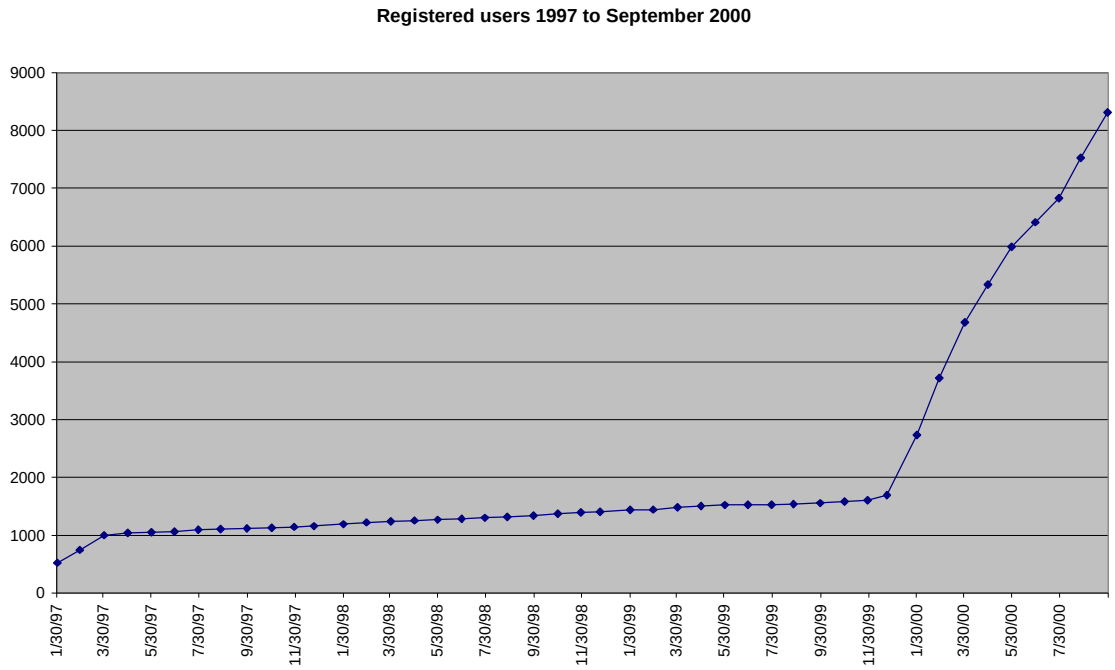


The number of users registered before 1 October 1999 was *1.555*, less than *20%* of the corresponding number one year later. ***Chart 1*** shows the *new users* registered each month in the four years of SSD. In the period 1997-1999, the monthly increases were modest except for the first 3 months of 1997 and the last month of 1999. For the 9 months of year 2000 included in this study, during which no fees were collected, the increases were significantly higher. Free access is likely to explain the increases in the first 3-4 months, just as the increases in the corresponding months in 1997. Increases in the later months of 2000 are likely to be the impact of the general explosion in computer equipment and Internet use. ***Table a*** (page 25) shows the details for each month from January 1997 to September 2000.

The development is even more clearly demonstrated in ***Chart 2*** with *cumulated* user numbers for the period. The increase in registered users from 1 December 1999 to 30 September 2000 is approximately linear indicating that in addition to the free access; the popularity of SSD is also affected by the general increase of people having Web access. The trend, which can be read from the chart, indicates that the number of users may pass *10.000* before the end of the year.

Departure of registered users cannot be deduced directly from the current system data. Indirectly, however, the activity of the different age cohorts by registration may give a basis for estimates of the departure of old users.

Chart 2:



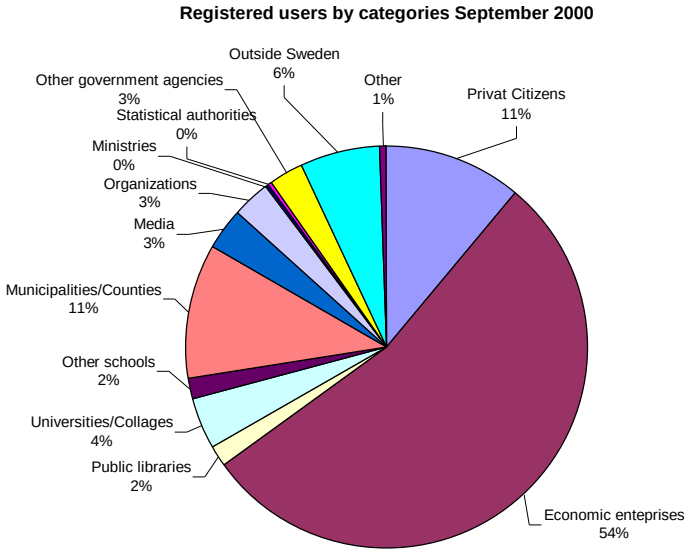
Which categories do the users represent?
The user registrations include a field called type or *category* of user. Thirteen user categories are distinguished. The information in this field is quite complete, but based on user applications for access to SSD. No control of the submitted information seems to have been implemented.

Chart 3 , on the following page, shows the relative distribution by categories of the *8.316* registered users. *Economic enterprises* account for *54%* of the users. *Private citizens* and *Municipalities/counties* share second place with about *11 %* each. Some users excluded from the registered user population because of invalid user-id may have belonged to *Municipalities/counties*. The traditional users of statistics, *Ministries and other government agencies*, are few among the registered users.

Table b permits comparison of the distribution by categories with the corresponding distribution at the end of September 1999. While in 1999, *Municipalities* were the leading user category accounting for almost *34 %* of the users; the distribution for September 2000 has changed significantly. Relatively, *Private citizens* increased from about *1%* to more than *10%*, *Economic enterprises* increased from 26% to 54% while *Municipalities went down from 34% to 11%* even though the absolute number of users increases with 369 users.

The interpretation of user category 12, *Outside Sweden*, represents a problem. Among the users who gave a foreign country address, *371* also answered Outside *Sweden* as their user category while *200* gave other user categories. On the other hand, *157* of the users, who did not give foreign country addresses, answered *Outside Sweden* as their user category. ***Table d*** shows details about the domestic and foreign users by user categories.

Chart 3:



Where are the users located?
The user population has a wide geographical distribution. The source for information about the users' national affiliation is the applications submitted by the users to obtain user identification and password. In the form, they were asked about their address, including country. We have assumed that those submitting foreign country addresses are located abroad.

In the user population of *8.316* at 30 September 2000, 571 users had country addresses other than Sweden. ***Chart 4*** shows the most common foreign addresses and their relative frequencies. The Nordic countries, USA, UK and Germany dominate as the location of SSD foreign users. In total, *50* country addresses, including Sweden, were represented among the registered users.

The trend from the previous year is interesting. At the end of September 1999, the number of registered user reporting a foreign country was *49*. While the total number of registered users increased with a factor 5,3, registered users with an address outside Sweden increased with a factor *11,7* indicating a significant growth in interest for statistical information from Sweden among foreign users.

Domestic geographic distribution can be illustrated using the Swedish Postal zip number. An imperfect, but indicative impression can be obtained by classifying the domestic registered users by the leftmost digit of the zip codes in 9 regions. ***Chart 5*** illustrates the relative distribution by regions and the foreign sector. ***Table e*** shows the users distributed by region at the end of September 2000 and the end of September 1999.

Chart 4:

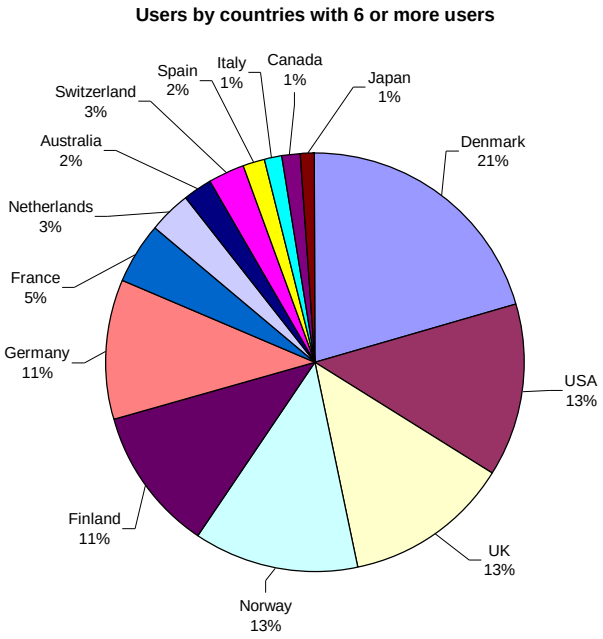
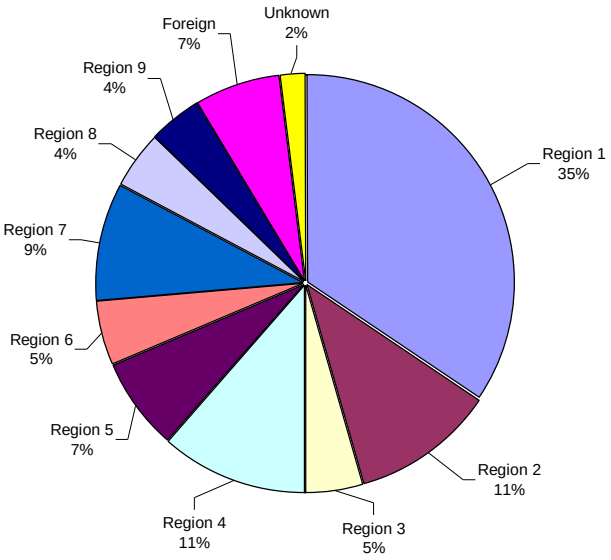


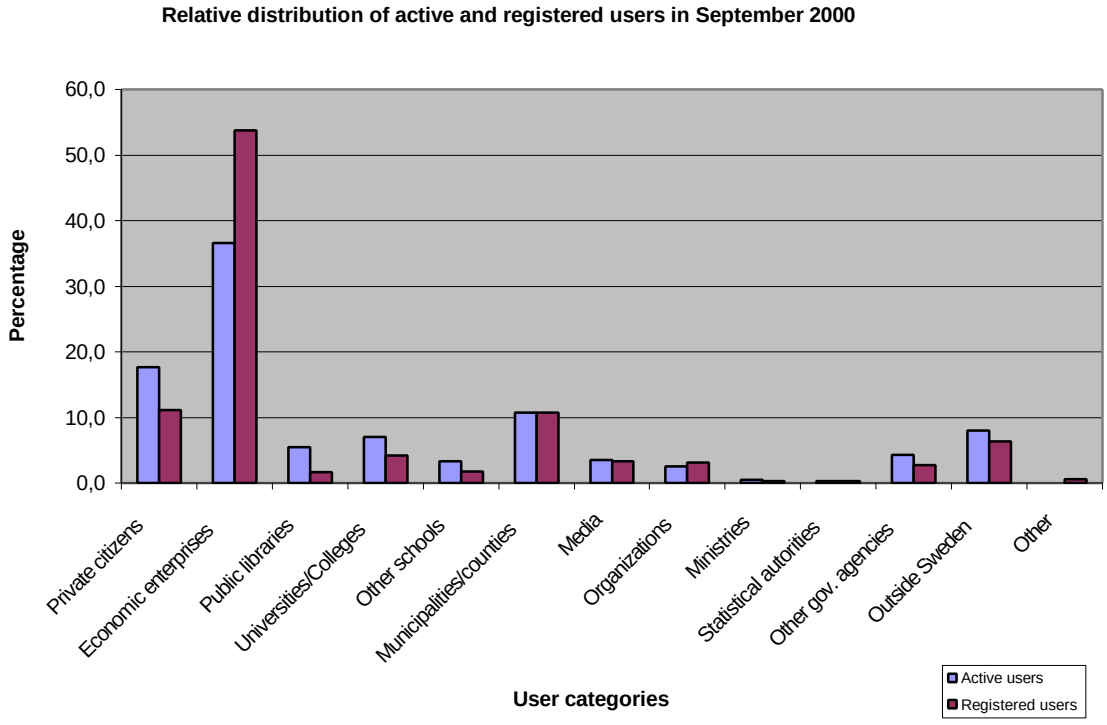
Chart 5:



How active are the registered users?
The SSD log showed that *600*, or about *7%*, of the *8.316* registered users accessed SSD one or more times during September 2000***.*** ***Chart 6*** demonstrates the relative distributions of the active users and the registered users by user category. From the chart we can see that particularly users in categories *Private citizens, Public libraries* and *Universities and colleges*, Other *schools* and *Other government agencies* were relatively more active than users in the remaining categories The categories for libraries and education have in common that they represent many individual users who may not be very active, but together they make the organizational user active.

The category with the largest number of registered users, *Economic enterprises*, was the dominant category of active users, but the percentage of active users from this category compared with the registered users was much lower than for the other categories. The explanation may be that each economic enterprise user represents a single need while in other categories, for example a library user, actually represents many independent individuals using the SSD facilities. Details of the distributions of active users compared with registered users can be read from ***Table f.***

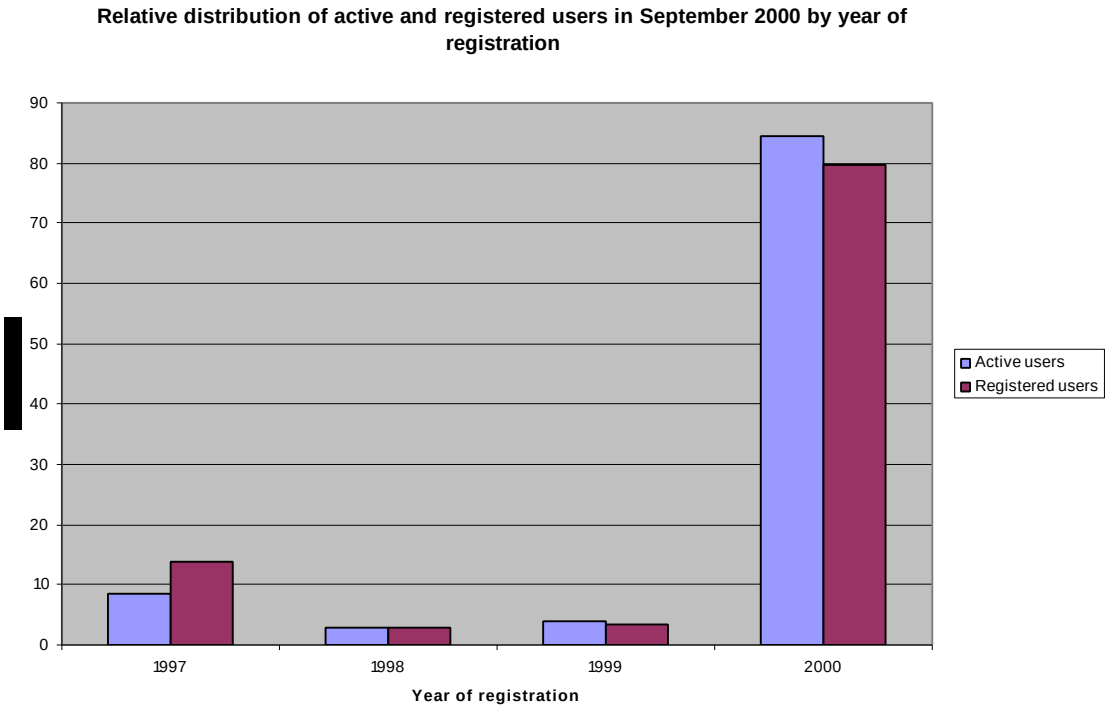
Chart 6:



Do the activities of registered users change over time?

Do users become less active after having had some time for exploring the databases? ***Chart 7*** displays the percentages of active and registered users in September 2000 by their year of registration. The chart shows that users registered in 1997 were relatively less active in September 2000, than the users registered in 2000, while for users registered in 1998 and 1999 there were no significant differences as to activities. ***Table g*** gives detailed figures.

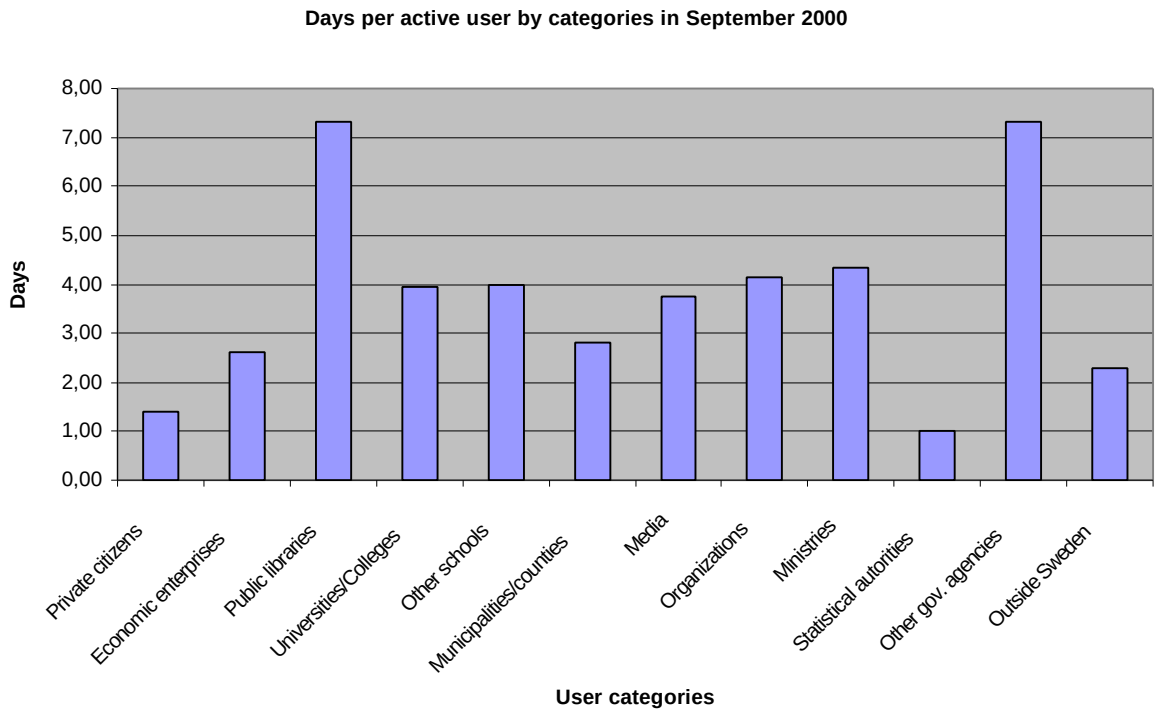
Chart 7:



How frequently do the users visit SSD?
Within the available data, we can raise 2 different questions. First, how many of the September 2000 users were active users for two or more days. Second, how many of the September 1999 users returned to make use of SSD in September 2000.

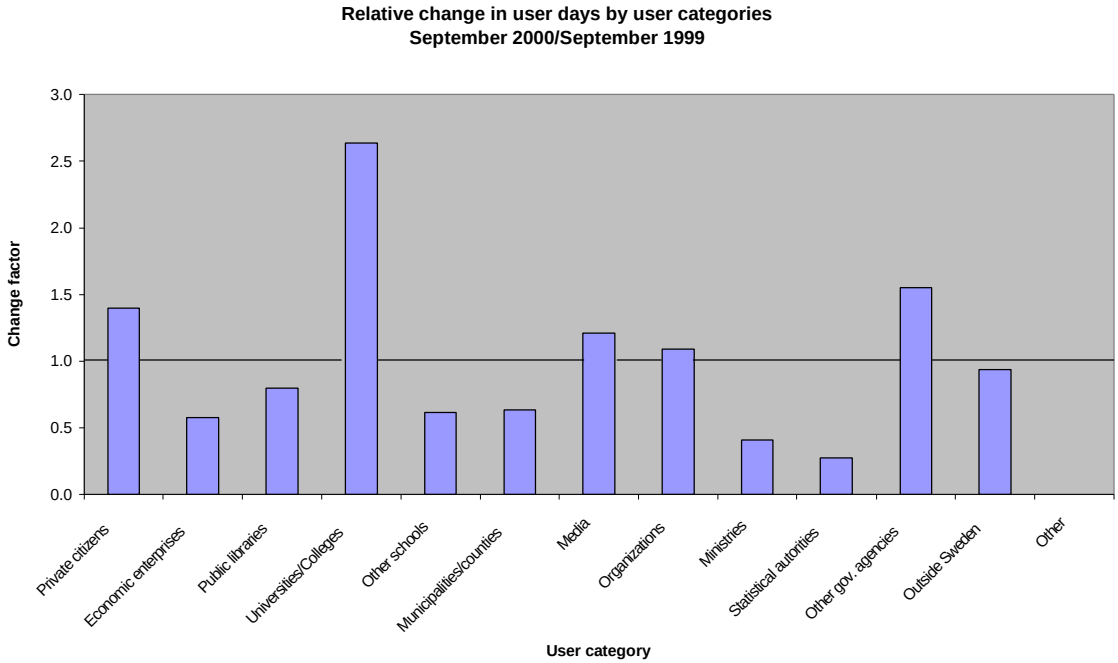
In this report, we use the term *session* to denote a user's activities within SSD during 1 day. The web-log shows that there were a total of *600* active users who had *1.841* sessions during September 2000, giving an average of *3,1* day sessions. ***Chart 8*** shows the average number of sessions per user in each of the 13 user categories. As might be expected, the categories with the highest averages were user categories 3 and 11, *Public libraries* and *Other government agencies*. Again the explanation is probably that each registered user in these categories has a number of individual users, with individual tasks for which they need statistical support.

Chart 8:



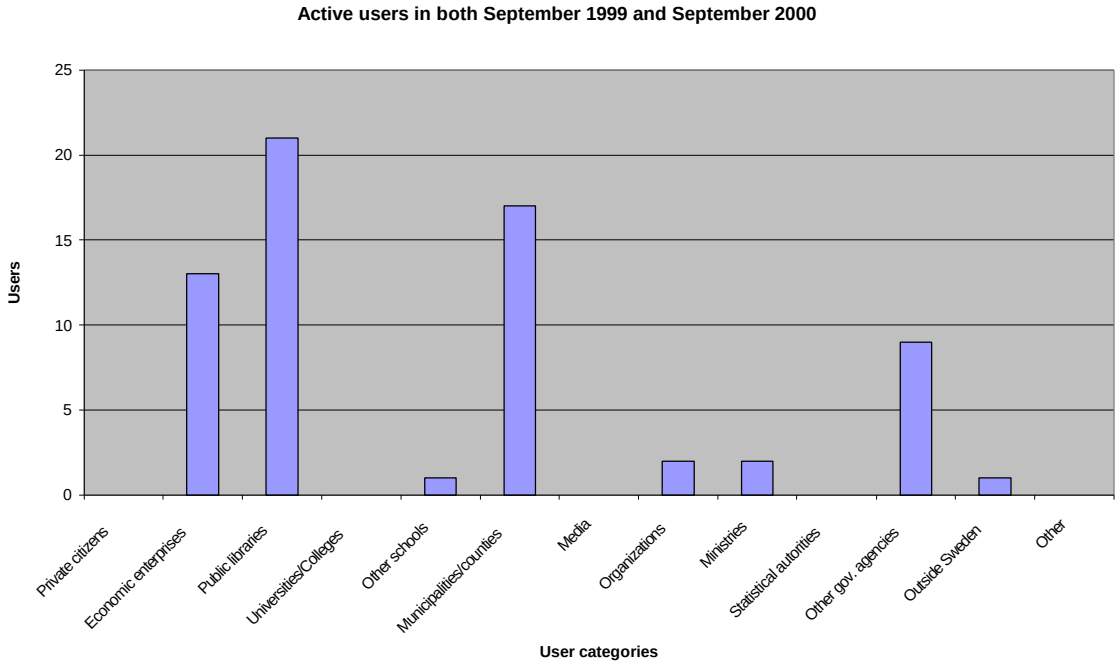
When we compare re-visits in September 1999 with September 2000, we observe a significant change. In September 1999, the *243* active users visited SSD in average 5.1 days. The decrease of average re-visits to *3,1* days in 2000, can be the impact of relative change in the distribution of users by categories. We have seen that the absolute number of users as increased much from 1999 to 2000, but the relative increase may have been greatest in categories with low tendencies to repeated use. In ***Chart 9***, the relative changes in sessions per user from 1999 to 2000 are displayed. Relative increases were particularly marked for the categories *Private citizens*, *Universities* and *Colleges* and *Other government agencies*, while the large category *Economic enterprises* reduced its sessions per user with almost *50%.*

Chart 9:



Users who had sessions in both September 1999 and September 2000 are considered long-term users. Out of the *246* active users in September 1999, *66* or about ¼ were also users in September 2000. ***Chart 10*** *shows* that the long-term re-visits were particularly strong in *Economic enterprises*, *Public libraries, Municipalities and Counties* and *Other Government Agencies*.

Chart 10:



We have commented on the relatively high usage from *Public libraries* and indicated that the explanation may be many secondary users. The 3 other categories with many revisits have users with permanent re-appearing tasks, which may explain their annual re-visits.

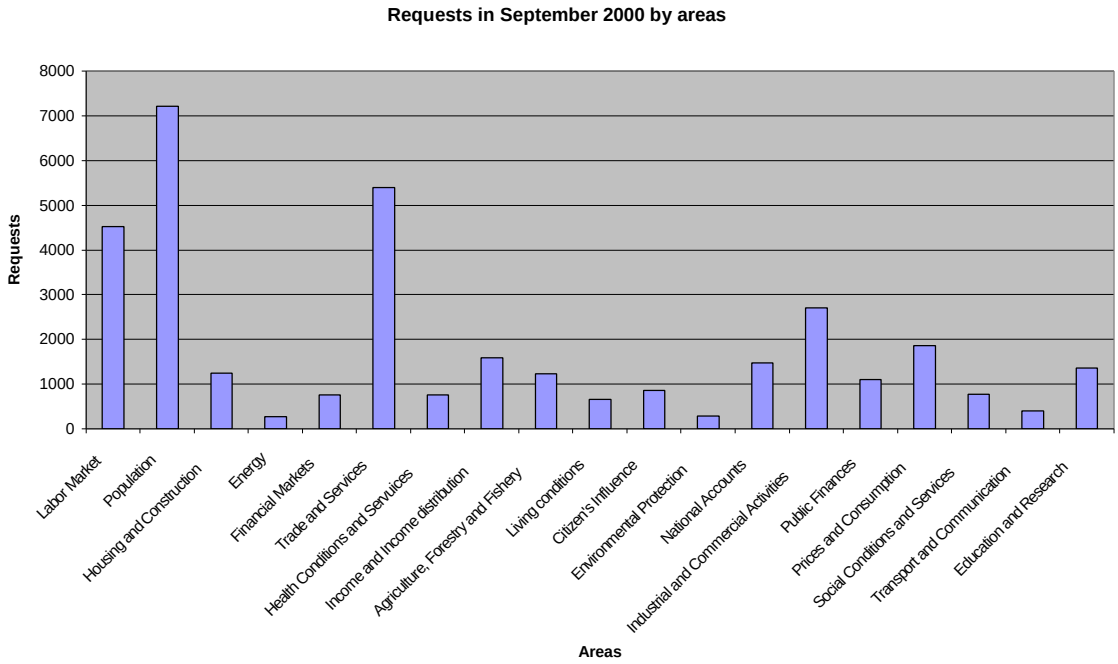
Which keywords do the users apply?
One of the options for locating desired statistical information in SSD is to search by *keywords*. In the keyword file for SSD, keywords are associated to the user, however there is no date or link to Web-log entries, making a usage study of individual keywords impossible. In September 2000, the users made *4.154* search requests. After elimination of some obviously meaningless records, the keyword file contained only 456 records. 2 different external users used all the keywords appearing in the keyword file. The data in the received keyword file were deemed unsuited for further analysis.

7. Requested and retrieved statistics

Which statistics are requested?
At the end of September 2000, external users could get access to and select among more than *800* statistical tables in SSD[1]. Users can search for statistics by either a hierarchy of *menus ('hovudtabelll', 'deltabell'*) or a *search* ('*sök* ') function. The highest menu level is here called statistical *area* ('*område'*). We have concentrated the analysis to classification of requests by the 19 statistical areas.

Of the *53.598* records in the web file for September 2000, *4.154* records were for the initiation of the search utility, *34.390* were related to the statistical areas, and the remaining *15.054* were output requests. ***Chart 11*** shows the distribution of the *34.390* search requests by statistical area. The illustration shows that requests are most frequently made to the areas *Population, Trade and Services,* and *Labor Market,* while the 2 least requested areas are *Energy* and *Environmental Protection.*

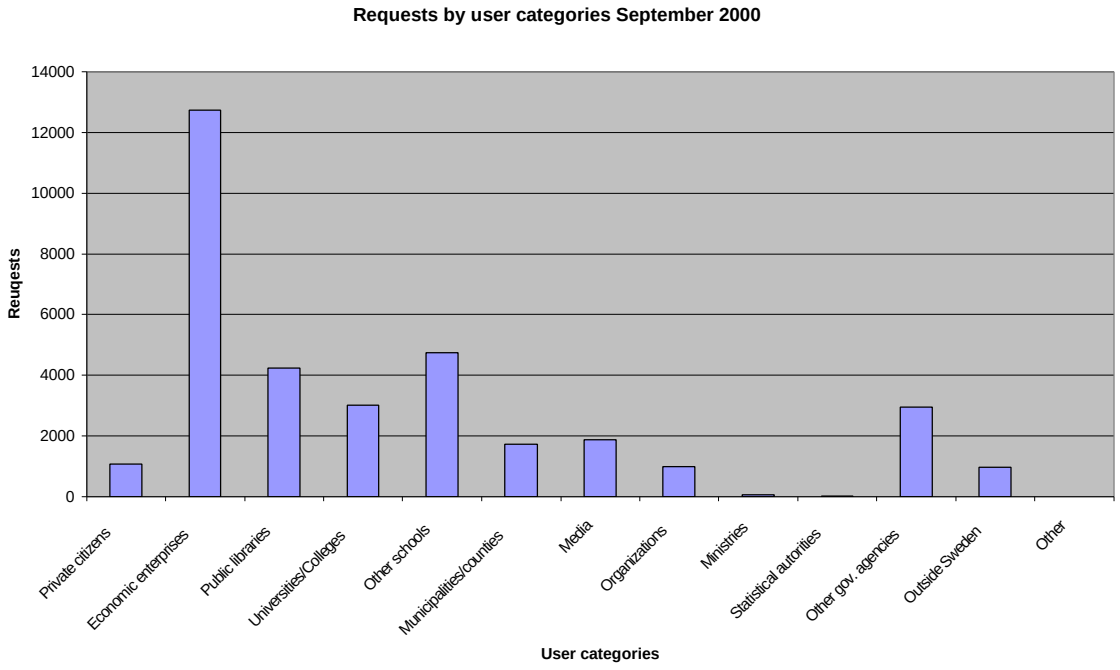
Chart 11:



[1] SCB: Lista över innehåldet i SSD.

What is the request frequency by user category?
In Chart ***12*** displays the same requests by user categories. Nearly 40% of the requests came from *Economic enterprises,* while the users in the smaller categories - *Libraries, universities, and other schools* - who represent *12%* of the registered users, request another *40%* of the total. One probable explanation as to why users from *Other schools* make so many requests is that there are many students behind each user identification. The same argument is probably also be valid for *Public libraries* and *Universities/Colleges*.

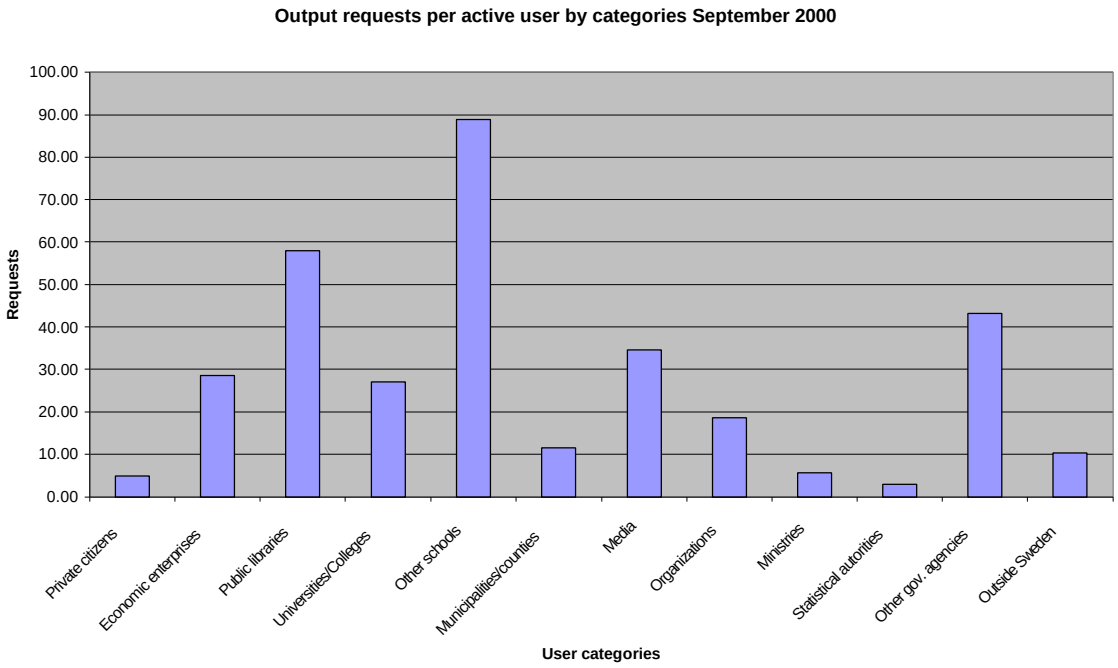
Chart 12:



Which statistics are requested by user category?
Table h gives the total requests for the areas cross-classified by user group. *Population statistics* were most frequently requested, accounting for 21% of all requests. Thereafter followed Trade & services (16%), Labor (13%), and Industrial & commerce (8%). There were only 13 requests from *Statistical authorities* (as external users). *Labor statistics* were most frequently requested by *Organizations* and *Ministries*, while *Trade statistics* were most frequently requested by *Other government agencies*. All other categories requested *population statistics* most frequently.

Chart 13 shows the distribution of output requests per active user. The chart indicates that the number of requests for output was highest from *Other schools* and *Public libraries*. As pointed out above, there might be a large number of secondary users behind each registered user in these categories. These secondary users might have had a short time each and have been more focused on direct output than other users. The figures also show that the educational system was becoming a major user of official statistics. *Media* and *Other government agencies* are two other heavy users.

Chart 13:



How are data retrieved?

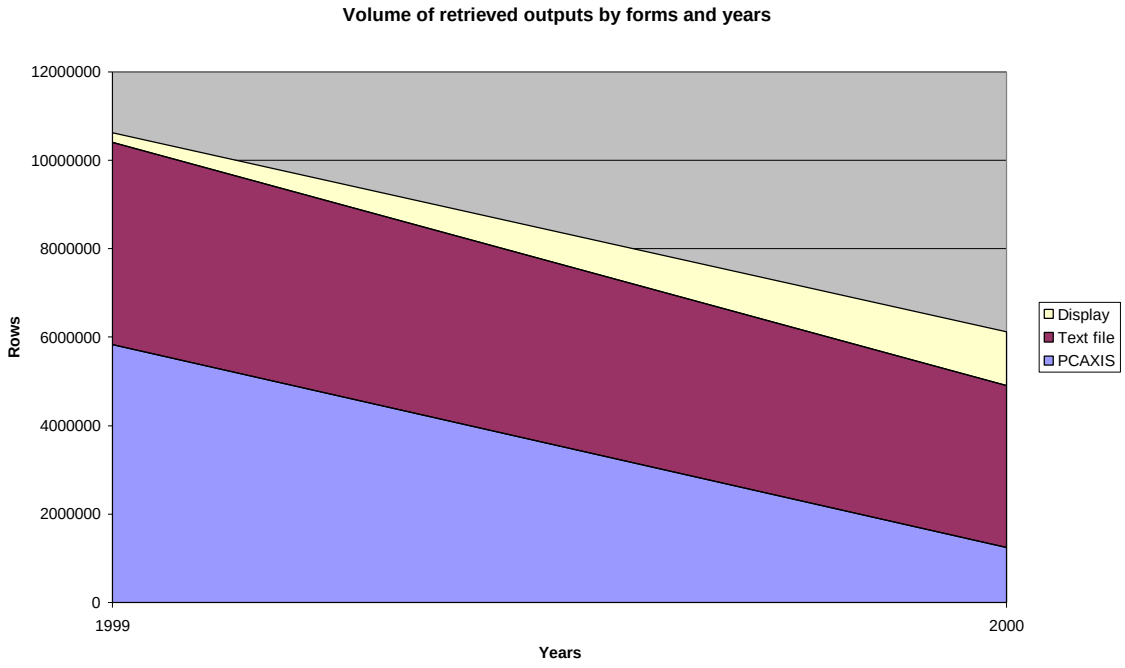
The *15.054* output requests returned statistical information in the form of a display on the user screen or a downloaded file. Screen displays constituted *2/3* of the output dissemination, while *1/3* returned files for the users' own processing. The users have *2* form options for files, ordinary text files or PCAXIS files for those users having acquired the SCB/PCAXIS software. The number of text files requested was nearly 3 times that of PCAXIS files*.*

While the total number of requests increased with a factor of 3.7 from September 1999 to September 2000, output requests doubled for the same two months. In other words, the amount of searching for statistical information increased significantly more than requests for output, probably reflecting an increase in 'explorative' users, for example *Private citizens*. ***Table i*** shows that the increase of output has been particularly strong for downloaded text files.

Table j shows details the number of output requests by category and requested output form. Display of statistics (*show*) on the user screens counts for *2/3* of all output requests. The users in *Universities and Colleges* and *Municipalities and Counties* downloaded what they found without previewing more frequently than the other categories. The explanation can be that these users were well acquainted with the statistics and the SSD and knew what they needed.

What is the volume of downloaded statistics?
While users who viewed statistics on their PC screens dominated the output requests, the picture became different when looking at the volume of the downloaded statistics defined as rows of output. ***Chart 14*** shows that there was a fall, from *10.628.037* to *6.126.903* rows with statistics downloaded to the users when we compare September 1999 with September 2000. The main reduction was due to fewer downloads of PCAXIS files. There was a certain increase in the volume of statistics displayed on user screens. Explanations for the reduction in the volume of output may be many, e.g. the users may have become more focused and knew what they wanted.

Chart 14:



8. Sessions in SSD

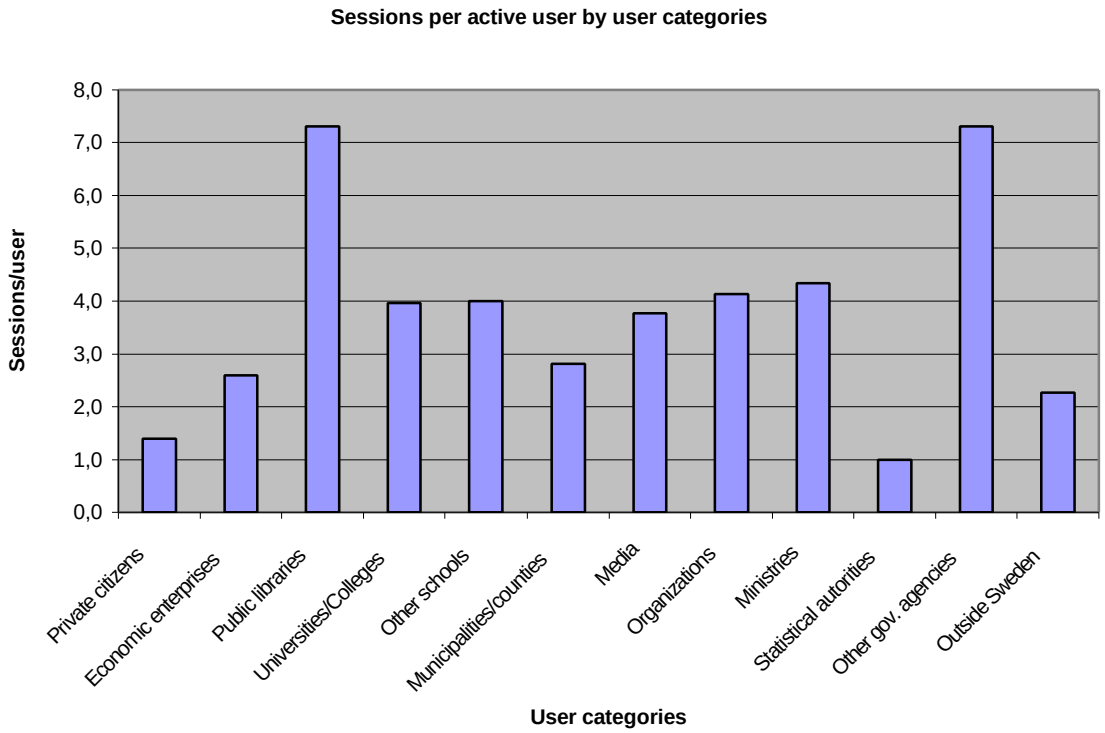
How many sessions were recorded in September 2000?
There were 600 different active users of SSD in September 2000. Ideally, we need a session concept that delimits a set of requests related to a particular task. In the above, we have used the concept a day session defined as all use of SSD by an identified user during a day. As noted earlier, this is due to an inability to separate concurrent secondary users using the same user-id. There were *1.842* day-sessions in September 2000 indicating that each active user used SSD in average *3,1* times during the month.

The number of requests in a session is interesting because it indicates the resources spent to find desired statistical information. It can also be a valuable indicator in evaluation of which information seems to be easy to find and which needs more resources and patience from the user.

Of the *1.842* sessions, *110* sessions included only one request. The remaining *1.732* sessions consisted of multiple requests, up to a maximum of *1.697* requests for 1 session, or about *3%* of all requests during the month. The single request sessions (*6%)* retrieved only meta-information about one statistical area. Most of these sessions probably represented users who wanted to see how the system functioned without any particular task as a background motive.

Chart 15 shows the sessions per active user in each user category. While the *Economic Enterprises* had most sessions, *569* sessions, it is the users in the categories *Public libraries* and *Other Government Agencies*, which had most sessions per active user. As pointed out in several sections, this must be interpreted with their many secondary users in mind. ***Table k*** presents the sessions users by user categories.

Chart 15:

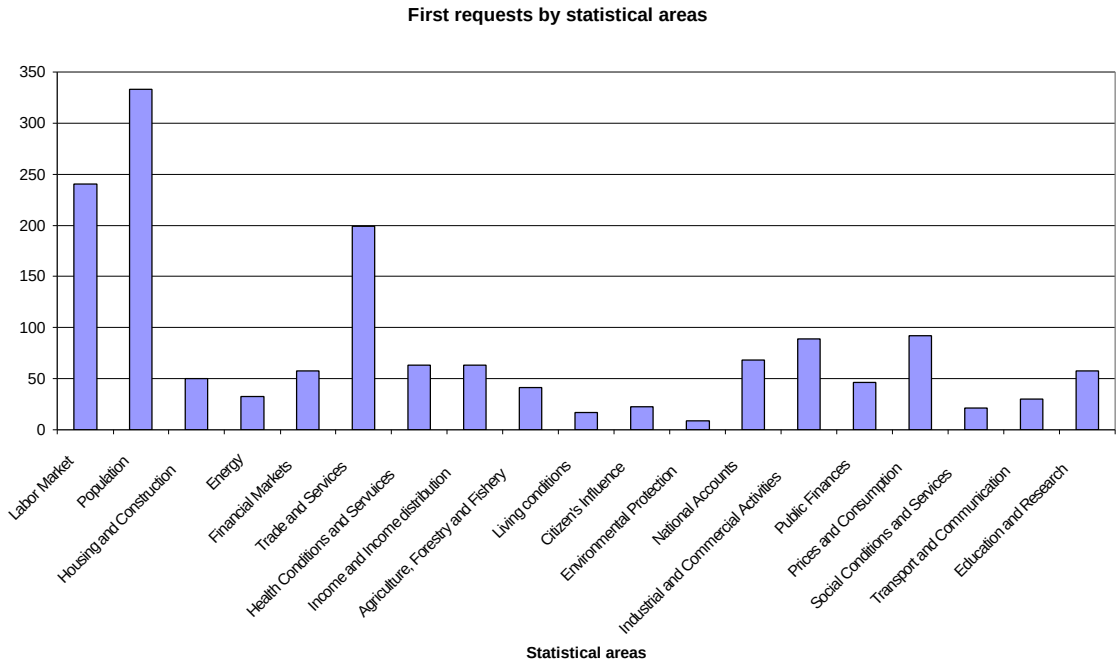


What was the most frequent start request in the sessions?
The behavior of the users when visiting SSD is an extremely important field to acquire knowledge about. Such knowledge may indicate needs, efficiency of the database design, and possibilities for improving the database.

A session can be started either with a search request (sök) or with a request for meta-information about one of the 19 statistical areas. In September 2000, there were *309* sessions starting with search requests and 1.529 starting with a request for a menu of a statistical area.

Chart 16 shows the session start requests by statistical area. The *Population* area was the most common start request with 330 sessions followed by *Labor Market* with 240 sessions

Chart 16:



The least frequently requested starting areas were *Environmental Protection, Living Conditions*, and *Social Conditions and Services*, all three represented only *1%* of the starting requests. Given the publicity these 3 areas receive, the indicated low interest from the statistical users is surprising. ***Table l*** gives details.

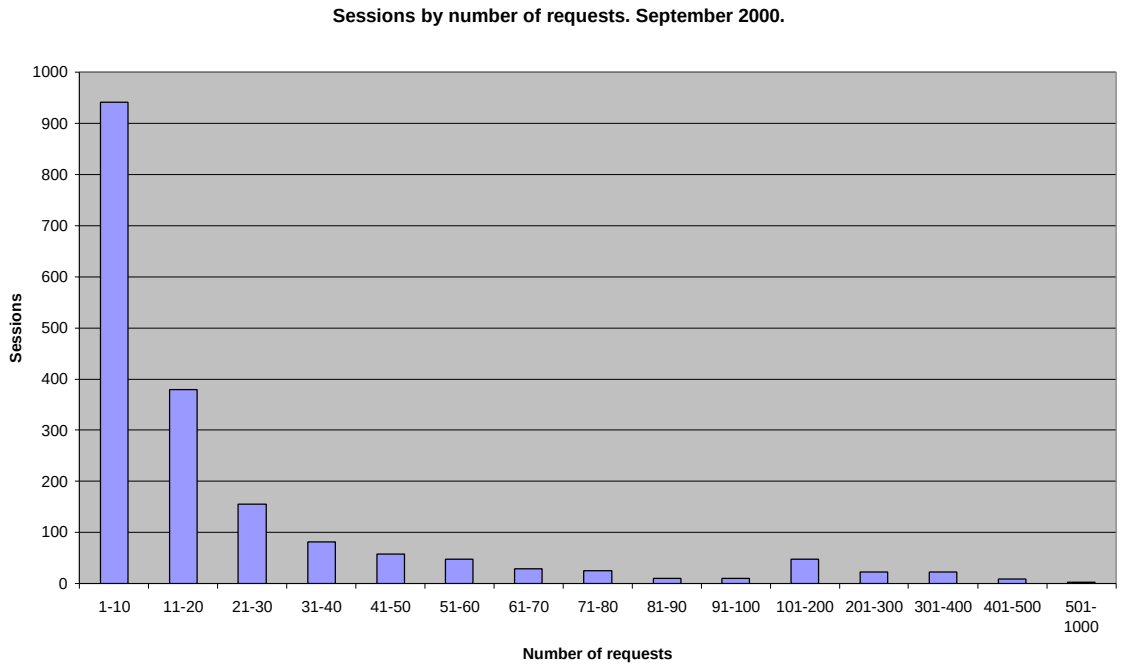
How many requests do the sessions include?

The second characteristic of a session is its extension or length. The length is a count of the number of requests a user has generated during one session (one single day). When counting requests, the count starts from the first request made after the entry page of SSD.

The average length of a session was *29* requests. Some of the users, *110* out of the *1.843* active users in September 2000, only made a single request and disappeared. On the other hand, the length of the longest session included is *761* requests. It should again be kept in mind that in some user categories, for example schools, there might be a number of secondary users using the same user identification. A long session can for example be the recording of students in a school class using the their school's or teacher's user identification.

Chart 17 shows the distribution of the sessions by length. ***Table m*** gives details of this size distribution. Note that the size group intervals are increasing in size.

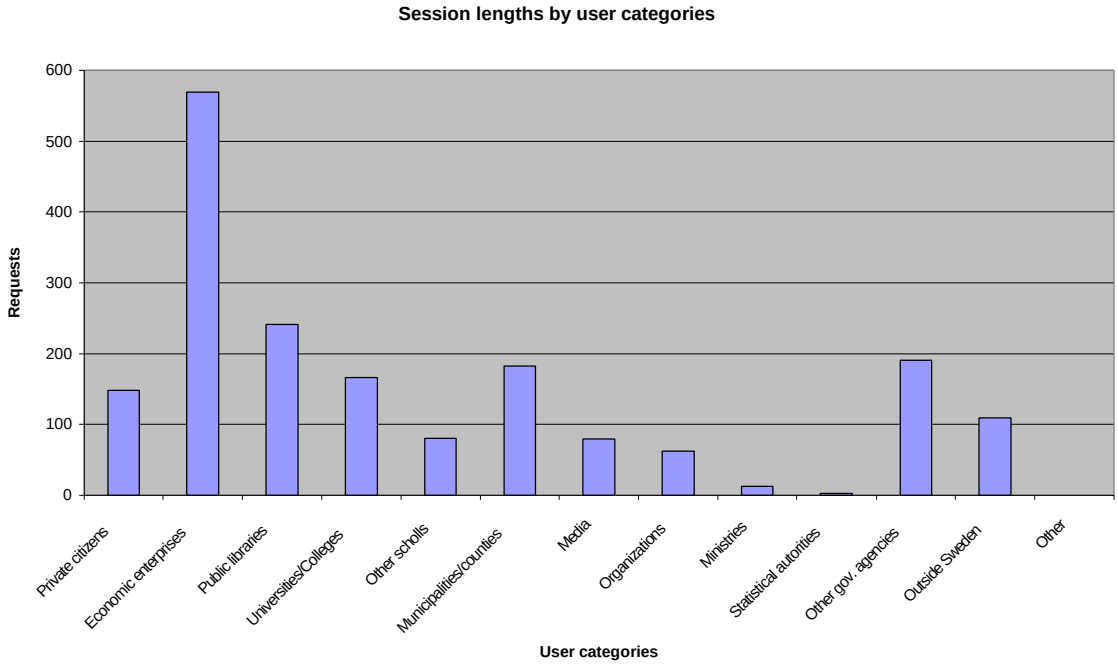
Chart 17:



How does session length differ by user category?

Given the size distribution of the sessions in the previous chart, it is interesting to inquire as to who is responsible for the very long sessions. ***Chart 18*** shows the average session length by user category.

Chart 18:

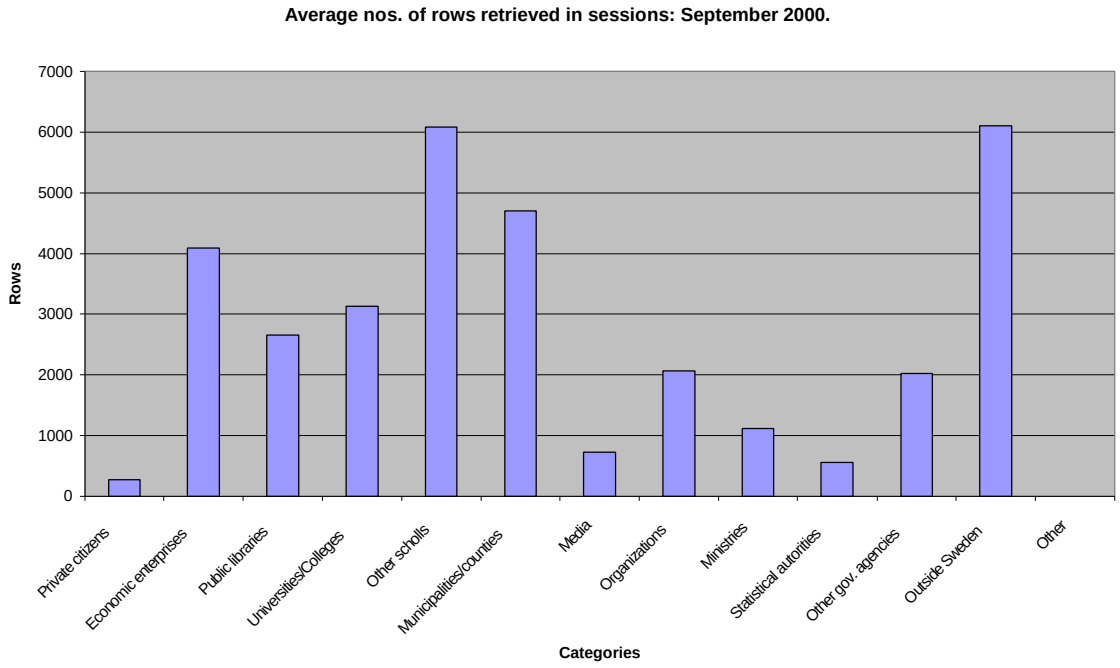


The chart shows that the *Economic enterprises* had the longest average session (even though the longest individual session is by a user in *Other schools*). *Public libraries* also rank high, but the explanation is probably many secondary users. ***Table n*** gives the statistical values.

What is the session row output?

The presumed aim of a session is to find statistics, which can either be displayed at the user monitor or downloaded, as files, for further processing. Section 8 above gives statistics on the volume of downloaded statistics per active user in September 2000. The average number of rows retrieved per session was *3.326* rows, which is a surprisingly large row number. ***Chart 19*** shows how the downloaded statistics, measured in rows, are distributed in sessions by user categories.

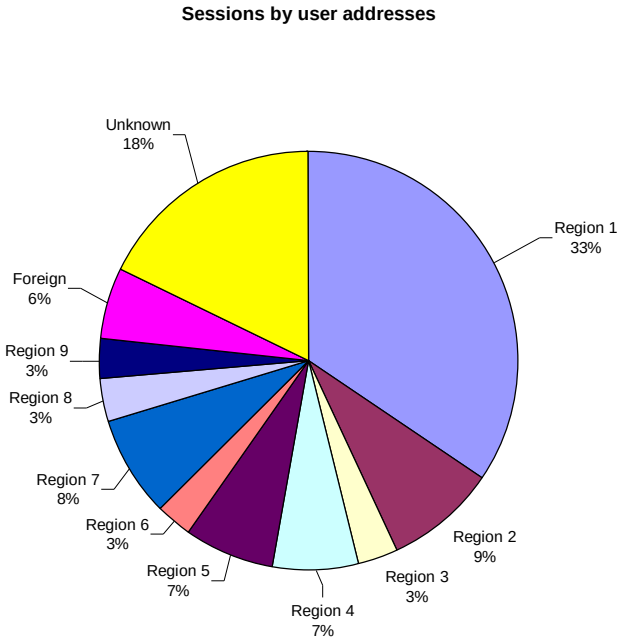
Chart 19:



We particularly notice the high volume of output per session from the user category *Outside Sweden*. So far in this analysis this category has not been notable, thus it is strange that it is such a significant category of users downloading statistics. As shown in *Table d,* this category of 371 users had submitted a foreign country address and 157 had also submitted domestic addresses or no address. One explanation can be that the users are foreign trade representatives, trade attachés in foreign diplomatic representations, etc. with a regular task to report statistics from Sweden.

From where are the sessions initiated?
The question about the geographical distribution of the users was discussed above. It is interesting to see if the sessions are distributed proportionally to the distribution of the users, or if the activities are concentrated to certain addresses***. Chart 20*** shows the relative distribution of the sessions, and ***Table p*** gives the details for the distribution.

Chart 20:



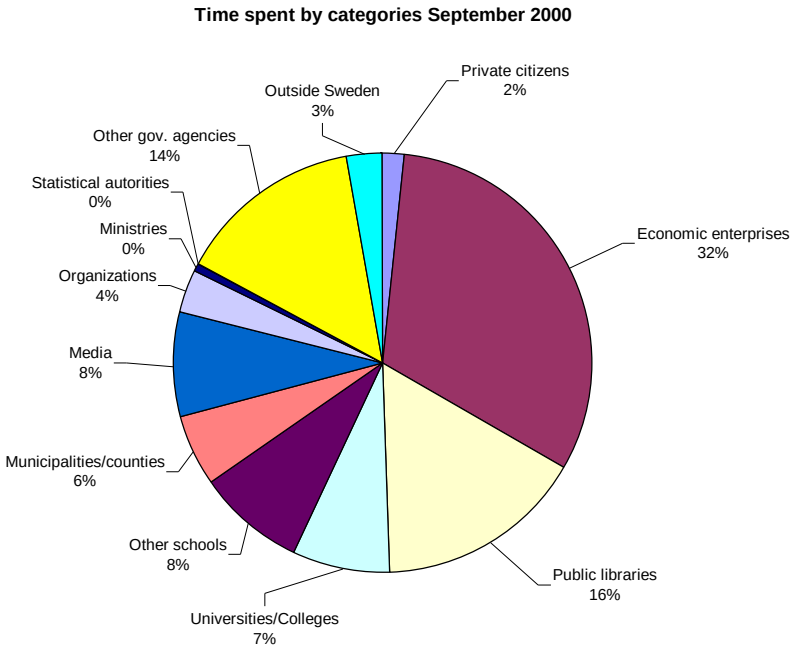
Comparing the pie chart above with chart 5 on page 8, the correspondence seems good with the exception that the group with users who have not given any address seems to be more active than the other groups. The users in the group *Unknown* are most probably the same as those who appeared in the category *Other* in ***Chart 19***.

9. Time spent in SSD

How much time did the users spend in SSD in September 2000?
Time spent in SSD is impossible to measure exactly. The log records the initiation of each request and total connect time can be calculated as the time between initiation of the first and the last request of a session. However, whether a user is working with a response from SSD, answering a telephone call, making a cup of coffee, etc. between *2* requests, cannot be logged. For this reason, any discussion of session time spent must keep these facts in mind.

The *600* active users spent *4.320* hours connected to SSD in September 2000, in average *7,2* hours each. The total hours were distributed on user categories as shown in ***Chart 21***. ***Table r*** gives details.

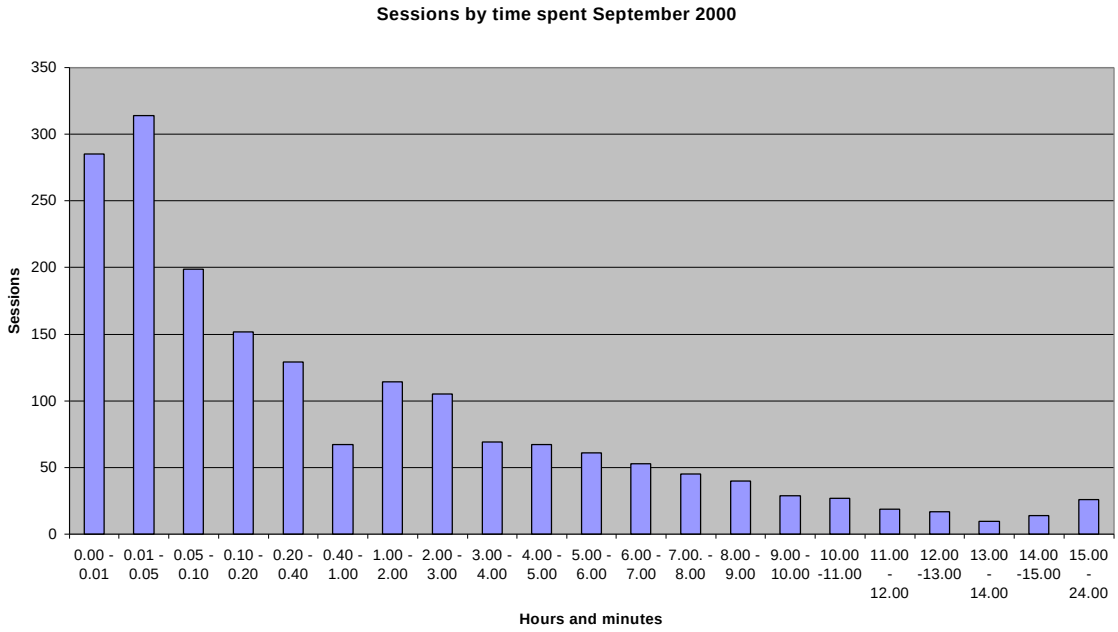
Chart 21:



How much time is spent on sessions?

The average number of connected hours per session was 2,3 hours. The time varies very much from *178* sessions logged with *0* time spent, because they only made a single search request and never returned again, to the session which lasted for nearly *24* hours. Session distribution by duration time is displayed in ***Chart 22.*** The details are given in ***Table r.***

Chart 22:



Compared with time spent with other types of databases, in which average time is measured in minutes rather than hours, a significant number the users spent long sessions with SSD. This may indicate that the interest for the statistical databases is based in more serious needs for information. A reservation must be made for the uncertainty in the delimitation of the sessions, which has already been discussed.

What is the time spent on requests?

There were *53.598* request made in September 2000 during the *4320* hours spent with SSD. The average time spent with a request was thus nearly *5* minutes, which is a long time compared with what is measured in other Internet sites. Also this figure is higher than found in similar studies in connection with other databases. These fact supports the idea that users of statistical databases seem to be more focused on the information wanted, and spend more time on studying the responses from the database.

10. Conclusions and recommendations

Conclusions

- User interest for SSD increased from September 1999 to September 2000 with a factor of *5.* Part of the increase was due to the 1 January 1999 abolition of the annual user fee. However, the increase started already in December 1998 and also continued after the first introductory months of 2000 indicating an additional impact from the general increase of computer and Internet use which most likely will continue also in the future.

- *Economic enterprises* were the dominate category of users, accounting for more than *50%* of all registered users at the end of September 2000. Users in this category were also among the most active. These enterprises have found in SSD an easily accessible source of statistical information for business decisions. As the SSD is further developed, it must be expected that this category of users will continue to increase and dominate the user population.

- Registered users were geographically widely distributed. SSD currently has only a partial support for foreign language users. Most of the meta-information used for navigation is still in Swedish. Still about *7%* of the users submitted foreign country addresses and *1/3* of these were from countries with Scandinavian languages. It would have been interesting to see which web sites the foreign users came from, e.g. if they came from web sites of other national statistical offices. When a completed English language version is implemented in SSD, the number of users from foreign countries will probably continue to increase.

- Not all registered users can be expected to be active every month. About *7%* of the registered users visited SSD one or several times in September 2000 while in September 1999, only *4%* of the registered users at that time visited SSD. *Economic enterprises* represented more than half of the active users. The users of this category have repetitive tasks each month, which require updated statistical data. This may be one of the reasons for the high activity of the users in this category.

- In September 2000, active users visited SSD, on average, *3,1* days. The corresponding average number of days for September 1999 was *5,1* days. The decrease in repeated visits can be explained with the cancellation of the annual fee, which attracted also users without economic motivation to exploit the database. The average number of user days per month must be expected to continue to decreas as the fraction of non-economically motivated users increases.

- *Public libraries* and *Other government agencies* were very active users in September 2000. Behind each registered user in these categories, there may be many secondary users, individual clients hidden for the SSD logging system.

- In comparison of September 1999 and September 2000 activities, *Universities and colleges* increased their activities relatively more than any other user category. This probably reflects the general strengthening of information technology in education.

- About *27%* of the active users on September 1999 were also active users of SSD in September 2000. This is a surprisingly high fraction and should be used with caution. If it gives a valid indication of long-term activities, which can be generalized, then about *200* of the active users in September 2000 will re-appear as users also in September 2001.

- The *600* active users in September 2000 made in total *53.598* requests to the database, of which *14.998* were requests for statistical output. The remaining were mainly requests returning meta-information to assist the users to identify the statistics in which they were interested.

- The search function (*sök*) was used *4.154* times. In the available data, recording of the search words and their relations to the sessions was unfortunately not implemented. This hindered an analysis of the importance of the search function in SSD.

- The average length of session was *29* requests including about *10* requests for statistical output. The session lengths varied from *1* single request up to 761 requests. The requested output also varied in volume with an average of *3.326* rows per output. This is a remarkably large average output. The users of the category *Other schools* had an average output two times the average, which indicates that many secondary users were active within each session (day).

- The users spent more than *4.300* hours working with SSD in September 2000. The user categories that spent most time on their SSD sessions were *Public libraries and Other government agencies*. The most probable explanation is that each of the users in these categories has many individual secondary users.

Recommendations

- References to statistical tables in the present log records are unsystematic and not well suited for analysis. The reference system should be reworked in such a way that when output is requested, a precise reference to the relevant table in the List of tables[2] is recorded. This would permit a more precise and meaningful classification of the data.

- In log systems, it is usual and easy to record the referring URL. This should also be included in the SSD log system. It would increase the possibility for more precise identification of the relations among records, and contribute to better analytical processing of the log.

- Some of the users, particularly in the categories *Public Libraries*, *Universities and Colleges* and *Other schools*, operate with many individual secondary users with separate individual objectives and tasks. Today's system does not permit identification of these users in separated sessions. Introduction of host IP in the logs can easily be introduced in the log system. This would help identify secondary users. To keep subsequent users on the same host apart is a second step. This could be solved by creation of a new temporary sub-identification each time the SSD opening page is requested.

- The search option is used in every fourth session. Important data were missing from the file in which the use of keywords should be recorded. It is recommended that the recording for this file be re-designed to include recording of each keyword used and reference to the resulting retrieved table as well as timing on when it was initiated and by whom. It would then be possible to include the search requests in an analysis of keyword uses.

- In the application form for SSD, user identification and password, the user categories include the option *Outside Sweden*. This alternative is ambivalent and does not belong in a user classification. We strongly recommend that this option be removed. In the present situation, a number of domestic applicants select this option, even though they have a Swedish country address resulting in a loss of information about their user category associations. Adding to the confusion, a number of users with foreign country addresses submitted user categories other than *Outside Sweden*. We recommend that this ambivalence be resolved by excluding this alternative from the user categories.

- The application form is extremely important as background information for interpreting the use of the database. We recommend that the form be re-designed with the analytical needs in mind as a

[2] Lista över innehållet i SSD.

distinct objective and that the user file record be revised according to the application form. As far as possible, registration of secondary users should be encouraged.

- We detected during the project work a number of duplicates, strange user identifications, etc. in the user file introduced during the manual registration of users. These problems seem to have disappeared when the automatic processing of user applications was introduced in the summer 2000. However, all possibilities for checking the completeness of the applications are not exploited in the registration process. There are still blank fields, strange zip numbers, unknown country names, etc. More can be done for automatic control of the submitted information.

- The present project used only a fraction of the available data in the user and log files. Much more can be done in many directions. Our recommendation, however, is that instead of spending resources on the present data with its many deficiencies, a new report system for periodic reports based on the proposed revisions of the user and log files be designed and implemented. It would give the staff responsible for further development of the SSD a valuable information source for decision support.

11. Tables

Table a:
Registered users by months. January 1997 to September 2000.

Months	*1997*	*1998*	*1999*	*2000*
January	525	34	30	1042
February	222	32	10	986
March	252	17	34	969
April	39	18	22	651
May	14	13	18	655
June	15	16	6	421
July	24	15	4	420
August	14	19	11	689
September	10	22	15	792
October	9	31	25	
November	19	20	19	
December	9	16	92	
Annual totals	1152	253	286	6625

Table b:
Registered users by categories. September 2000 and September 1999.

Category	*2000*	*1999*	*2000*	*1999*
Private citizens	928	14	11.2%	0.9%
Economic enterprises	4470	411	53.8%	26.4%
Public libraries	136	84	1.6%	5.4%
Universities/Colleges	349	105	4.2%	6.8%
Other schools	148	50	1.8%	3.2%
Municipalities/counties	894	525	10.8%	33.8%
Media	274	64	3.3%	4.1%
Organizations	257	53	3.1%	3.4%
Ministries	28	12	0.3%	0.8%
Statistical autorities	24	8	0.3%	0.5%
Other gov. agencies	230	99	2.8%	6.4%
Outside Sweden	528	123	6.3%	7.9%
Other	50	7	0.6%	0.5%
Total	8316	1555	100.0	100.0

Table c:
Registered users by foreign country addresses. September 2000.

Land	*Users*
Denmark	101
USA	66
UK	64
Norway	63
Finland	54
Germany	54
France	24
Netherlands	16
Australia	11
Switzerland	13
Spain	9
Italy	7
Canada	6
Japan	6
Belgium	5
Estonia	5
Iceland	5
Poland	5
Austria	4
South Korea	4
Hungary	3
Ireland	4
Latvia	3
Luxembourg	5
New Zealand	3
South Africa	3
China	2
Colombia	2
Greece	2
Litauen	2
Marocko	2
Philippines	2
Åland	1
Algerie	1
Brazil	1
Chile	1
Czeckia	1
Faroe Islands	1
Greenland	1
India	1
Israel	1
Kuwait	1
Malaysia	1
Mauritius	1
Mexico	1
Russia	1
Slovenia	1
Ukraina	1
Total	571

Table d:
Registered users by categories and foreign/domestic addresses. September 2000.

Categories	*Foreign users*	*Domestic users*	*Total*
Private citizens	92	836	928
Economic enterprises	65	4405	4470
Public libraries	1	135	136
Universities/Colleges	15	334	349
Other schools	2	146	148
Municipalities/counties	0	894	894
Media	7	267	274
Organizations	5	252	257
Ministries	5	23	28
Statistical autorities	4	20	24
Other gov. agencies	2	228	230
Outside Sweden	371	157	528
Other	2	48	50
Total	571	7745	8316

Table e:
Registered users by domestic regions. September 2000.

	Year	
Region	*2000*	*1999*
Region 1	2871	573
Region 2	911	160
Region 3	375	75
Region 4	942	156
Region 5	606	110
Region 6	410	83
Region 7	773	141
Region 8	365	70
Region 9	337	70
Foreign	571	50
Unknown	155	67
Total	8316	1555

Table f:
Registered and active users by categories. September 2000.

User category	*Registered*	*Active*
Private citizens	928	106
Economic enterprises	4470	219
Public libraries	136	33
Universities/Colleges	349	42
Other schools	148	20
Municipalities/counties	894	65
Media	274	21
Organizations	257	15
Ministries	28	3
Statistical autorities	24	2
Other gov. agencies	230	26
Outside Sweden	528	48
Other	50	0
Total	8316	600

Table g:
Registered and active users by years of registration. September 2000.

	Number of users		*Percentage*	
Reg. year	*Active*	*Registered*	*Active*	*Registered*
1997	51	1152	8.5%	13.9%
1998	18	253	3.0%	3.0%
1999	24	286	4.0%	3.4%
2000	507	6625	84.5%	79.7%
Total	600	8316	100	100

Table h:
Requests for meta-informationby user categories and statistical areras. September 2000.

Area	*User categories*										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Labor Market	92	1553	592	290	648	222	165	292	32	0	506
Population	285	2729	659	511	1275	623	447	102	7	0	532
Housing and Constructic	20	577	83	64	176	42	101	4	0	0	151
Energy	9	106	32	23	32	6	20	12	1	0	17
Financial Markets	25	406	117	42	83	8	24	7	0	0	27
Trade and Services	83	1954	481	354	499	252	320	249	0	8	699
Health Conditions and S	60	205	128	98	121	57	31	7	0	0	40
Income and Income distı	76	586	145	79	278	146	106	47	3	0	90
Agriculture, Forestry and	44	269	130	437	71	12	131	26	0	0	95
Living conditions	30	208	106	69	100	17	40	6	1	0	64
Citizen's Influence	21	73	469	127	104	37	2	4	0	0	23
Environmental Protectio	18	79	25	42	44	19	12	5	0	0	27
National Accounts	93	588	206	97	188	24	56	33	1	5	105
Industrial and Commerc	77	1336	284	169	256	47	177	80	0	0	220
Public Finances	18	410	217	111	158	31	80	25	0	0	48
Prices and Consumption	48	975	248	74	164	17	63	80	17	0	121
Social Conditions and S	38	145	66	44	317	71	32	3	0	0	51
Transport and Communi	12	135	82	32	63	14	16	2	1	0	41
Education and Research	24	392	175	340	178	83	41	8	1	0	94
Total	1073	12726	4245	3003	4755	1728	1864	992	64	13	2951

Table i:
Output requests by forms. September 1999 and September 2000.

Presform	*1999*	*2000*	*2000/1999*
pcax	656	1287	1.96
save	1534	3533	2.30
show	5188	10178	1.96
Total	7378	14998	2.03

Table j:
Output requests by forms and user categories. September 2000.

Categories	*pcaxis*	*save*	*show*	*Total*
Private citizens	42	43	430	515
Economic enterprises	363	1549	4338	6250
Public libraries	130	528	1257	1915
Universities/Colleges	55	400	684	1139
Other schools	245	329	1203	1777
Municipalities/counties	126	235	391	752
Media	32	70	624	726
Organizations	32	32	216	280
Ministries	5	1	11	17
Statistical autorities	3		3	6
Other gov. agencies	184	218	720	1122
Outside Sweden	70	128	301	499
Total	1287	3533	10178	14998

Table k:
Sessions, active and registered users by user categories. September 2000.

Categories	*Sessions*	*Act.users*	*Reg.users*	*Session/Act.user*	*Session/Reg.user*
Private citizens	148	106	928	1.4	0.2
Economic enterprises	569	219	4470	2.6	0.1
Public libraries	241	33	136	7.3	1.8
Universities/Colleges	166	42	349	4.0	0.5
Other schools	80	20	148	4.0	0.5
Municipalities/counties	183	65	894	2.8	0.2
Media	79	21	274	3.8	0.3
Organizations	62	15	257	4.1	0.2
Ministries	13	3	28	4.3	0.5
Statistical autorities	2	2	24	1.0	0.1
Other gov. agencies	190	26	230	7.3	0.8
Outside Sweden	109	48	528	2.3	0.2
Other	0	0	50	0.0	0.0
Total	1842	600	8316	0.2	3.1

Table l:
First requests in sessions by statistical areas. September 2000.

Areas	*Requests*
Labor Market	240
Population	333
Housing and Constructic	50
Energy	32
Financial Markets	57
Trade and Services	199
Health Conditions and S	63
Income and Income distı	63
Agriculture, Forestry and	41
Living conditions	17
Citizen's Influence	22
Environmental Protectioı	9
National Accounts	68
Industrial and Commercı	89
Public Finances	46
Prices and Consumptior	92
Social Conditions and Sı	21
Transport and Communi	30
Education and Researcl	57
Total	1529
Search requests	309

Table m:
Sessions by size groups. September 2000.

Requests in sessions	*Number of sessions*
1-10	942
11-20	380
21-30	155
31-40	82
41-50	58
51-60	47
61-70	29
71-80	26
81-90	10
91-100	10
101-200	47
201-300	22
301-400	22
401-500	9
501-1000	3
Total	1842

Table n:
Sessions by categories. September 2000

Categories	*Sessions*
Private citizens	148
Economic enterprises	569
Public libraries	241
Universities/Colleges	166
Other schools	80
Municipalities/counties	183
Media	79
Organizations	62
Ministries	13
Statistical autorities	2
Other gov. agencies	190
Outside Sweden	109
Other	0
Total	1842

Table o:
Average row outputs by categories. September 2000.

Categories	*Rows*
Private citizens	265
Economic enterprises	4092
Public libraries	2657
Universities/Colleges	3133
Other schools	6084
Municipalities/counties	4705
Media	722
Organizations	2061
Ministries	1121
Statistical autorities	560
Other gov. agencies	2023
Outside Sweden	6104
Other	0
All	3326

Table p:
Sessions by user addresses. September 2000.

Region	*Sessions*
Region 1	634
Region 2	160
Region 3	57
Region 4	122
Region 5	129
Region 6	48
Region 7	144
Region 8	60
Region 9	58
Foreign	105
Unknown	325
Total	1842

Table q:
Time spent by user categories. September 2000.

User cat.	*24 hour days*
Private citizens	3
Economic enterprises	57
Public libraries	29
Universities/Colleges	13
Other schools	15
Municipalities/counties	10
Media	14
Organizations	6
Ministries	1
Statistical autorities	0
Other gov. agencies	26
Outside Sweden	5
Total	180

Table r:
Sessions by time. September 2000.

Hours.min	*Sessions*
0.00 - 0.01	285
0.01 - 0.05	314
0.05 - 0.10	199
0.10 - 0.20	152
0.20 - 0.40	129
0.40 - 1.00	67
1.00 - 2.00	114
2.00 - 3.00	105
3.00 - 4.00	69
4.00 - 5.00	67
5.00 - 6.00	61
6.00 - 7.00	53
7.00. - 8.00	45
8.00 - 9.00	40
9.00 - 10.00	29
10.00 -11.00	27
11.00 - 12.00	19
12.00 -13.00	17
13.00 - 14.00	10
14.00 -15.00	14
15.00 - 24.00	26
Total	1842

Appendices

Appendix A

Data received

All files received from SCB were in. text format for easy conversion to other formats required in processing.
The users were recorded in 2 files, Internetkund*_personar.txt* and *Internetkund_person_00_09_01_09_30.txt.* The first is a list of all registered users of SSD from January 1 1997 to September 5 2000. They were combined to:

File: Internetkund_personar.txt

User-id
Kundtyp
Pnr
Startdatum
(Country)

The uses of SSD in September 1999 and September 2000 are recorded in 4 files, each month distributed about evenly between 2 files.

Files: Web_logfil_99_09_01_09_15.txt
Web_logfil_99_09_16_09_30.txt
Web_logfil_00_09_01_09_14.txt
Web_logfil_00_09_15_09_30.txt

User-id
Hovudtabell
Deltabell
Htmlsida
Datum
Presform
Antalrader
Antallinnehall

The 4 files were concatenated to one file for September 1999 and a second for September 2000.

A file with information about the keywords used by the different users in their searches was also received:

File: Web_sord.txt

User-id
Sokord
Alternativ

Since the records lacked time references, and only contained 4 different user-ids, the file could not be used.

Another file for September 2000 was also received. It was used for transferring the country addresses to and supplementing user records where needed in the *Internetkund_personar.txt*

File: Internetkund_person_00_09_01_09_30.txt

User-id
Abonnemang
Typid
KundNr
LopNr
Enamn
Fnamn
OrgNr
Fadr1
Fadr2
PNr
Ort
Adr1
Adr2
Ort2
PNr2
Land
Tele
Email
FaxNr
SDBKund
FriKund
KundType
KundNamn
KortNamn
Land2
KomKod
Logonkod
Datum
LevEmail

Appendix B

TERMINOLOGY

Active user See user

Area See Statistical area.

Category See User category.

Domestic user See user

External user See user

Foreign user See user

Observation periods
Two periods of observations were selected from the continuous logs for requests to the database, September 1999 and September 2000. The month of September was considered to be 'representative', gave a possibility to compare the development with one year's interval, and was the most recent month when the decision was taken.

Output request
An output request is characterized by a topic and a presentation form of 'show', 'save' or 'pcaxis'. The presentation form *('presforms')* indicates that statistical facts were presented on the users monitor, sent as a text file, or as a file that can be used by SCB's special program *pcaxis* for further processing.

Output volume
Number of rows with statistics in response to an output request

Population
The population, within which this study has been carried out, was the registered, external users at the end of September 2000. In determination of the study population, the source file of registered users was examined and duplicates, incomplete records, etc. were eliminated.

Registered user See user

Request
When a user is admitted to the databases, he may make a request based on a menu or list, or by means of keywords. A request response may be another menu, or a database process that later may result in output to the user.

Request time
The logged time between initiation of 2 requests, measured in number of hours and minutes.

Search request
A search request is characterized by a topic and the presentation form *('presform')* 'VISA', indicating that information, typically metadata about the topic, is returned to the user.

Session
A session is defined as the sequence of requests a particular user makes to SSD during a single day. The justification for this concept is that requests in a session probably are related to the same task. A session can therefore be a meaningful statistical unit for analyzing the variation of the recorded behavior of the users as well as repeated use by the same user. Sessions are composed of a varying number of requests, different initial requests, and different requests in the sequence of requests.

Session Length
Number of requests recorded in a session.

Session Time
The logged time between initiation of the 1st and last request from a unique user during one session, measured in number of hours and minutes.

Secondary user See user

Statistical areas
The statistical tables included in SSD are grouped in 19 statistical areas:

AM	Labor Market
BE	Population
BO	Housing and Construction
EN	Energy
FM	Financial Markets
HA	Trade and Services
HS	Health Conditions and Services
IF	Income and Income Distribution
JO	Agriculture, Forestry and Fishery
LE	Living Conditions
ME	Citizen's Influence
MI	Environmental Protection
NR	National Accounts
NV	Industrial and Commercial Activities
OE	Public Finances
PR	Prices and Consumption
SO	Social Conditions and Services
TK	Transport and Communication
UF	Education and Research

User

In this report, users are distinguished between: Active, Domestic, External, Foreign, Registered, and Secondary users, as defined below.

Active user is a *registered* user using the SSD during September 1999 or/and September 2000. An active user in a period may have one or more sessions during the period.

Domestic user is one who has not submitted a foreign country address when applying for access to SSD. There may be users who have given local foreign addresses, but omitted the country address. Since there is no editing of the address information, these were counted as domestic users. Each domestic user is assigned to a region based on the leftmost digit of zip codes submitted in the domestic addresses.

External user is a user, who does not work in or for SCB or has been given access by some special arrangement. In this study, it was decided to include only external users.

Foreign user is one who has submitted a foreign country address. Each foreign user is assigned to a country based on the country address.

Registered user is defined as a person or organization registered in SSD with a unique user identification in the period 1 January 1997 to 30 September 2000. A single physical person or organization may have several user identifications and thus be counted several times, e.g. several employees in a company working in parallel on the same task. On the other hand, several physical persons and several different tasks may be hidden behind single user identification. Examples include a library using single user identification on behalf of several clients. Each registered user is assigned a registration date.

Secondary user is a client or staff member of a primary/registered user who has access to the primary user's unique identification code. Some of the registered users such as libraries, schools, large agencies, etc. may make their equipment and there access to SSD available to a number of *secondary users* such as library visitors, students and agency staff members.

User category

Based on the application for access, each user is assigned to one of 13 different user categories:

1. Privet citizen
2. Economic enterprise
3. Public library
4. University/college
5. Other school
6. Municipality/county
7. Media
8. Organization
9. Ministry
10. Statistical authority
11. Other government agency
12. Outside Sweden
13. Other

Appendix C

SN, 3/8-00

Log file for 16.april 2000 - An illustrative pilot study.

Received log files

Two log files were received from SCB in July. Both contained records for April 16, 2000. One file with 4,5 Mb was in 'extended windows' format, the other with 4 Mb was in NCSA format.
The first format is most convenient for this project and has been studied. Most of the information we needed was found in the files.

Some concepts

We refer to the SCB equipment delivering information as the ***host*** and equipment requesting information as a ***client.*** Each client connected to the host is identified by an ***IP-number***. A request from a client is resulting in a ***page*** delivered from the host. Each page is frequently composed of several ***components***, each of which is recorded in the log. The log will therefore contain a number of lines for each page. A chain of requests from a client by links from one page to another is referred to as a ***visit***. The person operating the client is called a ***visitor***.

Some clients have a permanent IP-number, while others are assigned IP-number dynamically by their net provider each time they connect to the net. It is therefore not possible to determine the exact frequency of re-visits by each client during a time interval.

We will use the visit as the primary ***statistical unit***. To distinguish 2 or more visits from the same client, the time between two consecutive requests from the client must be within a preset length. Such a restriction was not used in this pilot study.

Most clients will save received pages for some time in a ***cache***. Repeated requests for a page during this time will not be sent to the host and therefore not recorded in the log. The number of pages recorded in the log will therefore be a low estimate of the number of pages seen by the client users.

Questions raised

Lars Nordbäck listed some points of interest for SCB in an e-mail from June 21:

Kom kunden till databaserna direkt eller gick man via andra länkar. T ex från någonstans på vår egen webbplats eller utifrån, eller via sina favoriter: Referring URL is not logged in the files received. It was therefore not possible to analyze from where the visitor came if she came from outside SCB. *It is possible to* follow the paths within SCB.

Finns det något mönster i vart användarna tar vägen: It is possible to reconstruct the visit path within SCB as far as all requested URLs are logged, but it is not possible to find where the users go.
Gör man under en session både en tabell på skärmen och sedan laddar ned samma material eller görs det raka vägen till en nedladdning. Är det någotmönster i att titta och gå till textfiler eller till PC-AXIS-filer? It should be possible to distinguish between the two approaches and identify if the visitor requests text files and/or PC-AXIS files.

Är det något beteendemönster kopplat till vissa material? Yes, we consider the identification of such

patterns to be a central part of the project.

Hur stora material tar kunderna fram, finns det något mönster i det? För olika statistikområden? The log should in principle reveal how many incidents of downloading occur in each visit and also the sizes of files downloaded. Whether it will be possible in practice is not certain.

Söker man bara 1 material i en session eller är det vanligt att ta frånflera material vid samma session: This should be possible to study.

Kan man göra någon koppling mellan internet-loggarna och de loggar vi hari databaserna: From the exhibit log, we can read the access to the databases, but we do not know enough to conclude if the 2 types of log files can be matched. It would be interesting but probably require some efforts.

Kan man se om samma användare dyker upp vid alla tre månaderna och tarsamma material: As indicated above, clients with permanent IP-numbers can be identified from day to day, but those with dynamically assigned IP-numbers cannot.

Är det några som kommer mycket frekvent, varje dag eller varje vecka: The logs should permit this kind of study, but because of the size of the logs an accurate study will require a lot of resources. Estimates of frequencies should be possible to develop.

In addition to these questions, we imagine that the time spent at different pages, pattern of visit times by different categories of clients, etc. would be of interest.

A pilot study of the exhibit file for 16.april 2000

The date April 16, 2000 was Palm Sunday and not a representative weekday. The file must therefore be regarded as an exhibit of a log file and the following figures as illustrations of information, which can be retrieved from log files.

Length of observation: 24 hours.
Number of visits: 1.427 (no time restriction between requests).
Logged component records: 38.639
Number of different components: 4.671.
Number of pages: 21.274 (records with .gif, /images/, /nytt.gif, /pil.gif, /press/ eliminated)
Number of different pages: 4.479.

Visits per hour: 59.
Pages per visit: 15.

Visits per page: 0.31.

Figure 1 shows the distribution of visits by number of pages requested. It indicates that 241 of the visitors request a single page. Of the 1.400 visitors, 522 visitors requested between 11 and 39 pages. A more detailed inspection revealed that one visitor downloaded 6.520 pages meaning that the visitor repeated requests for many pages several times.

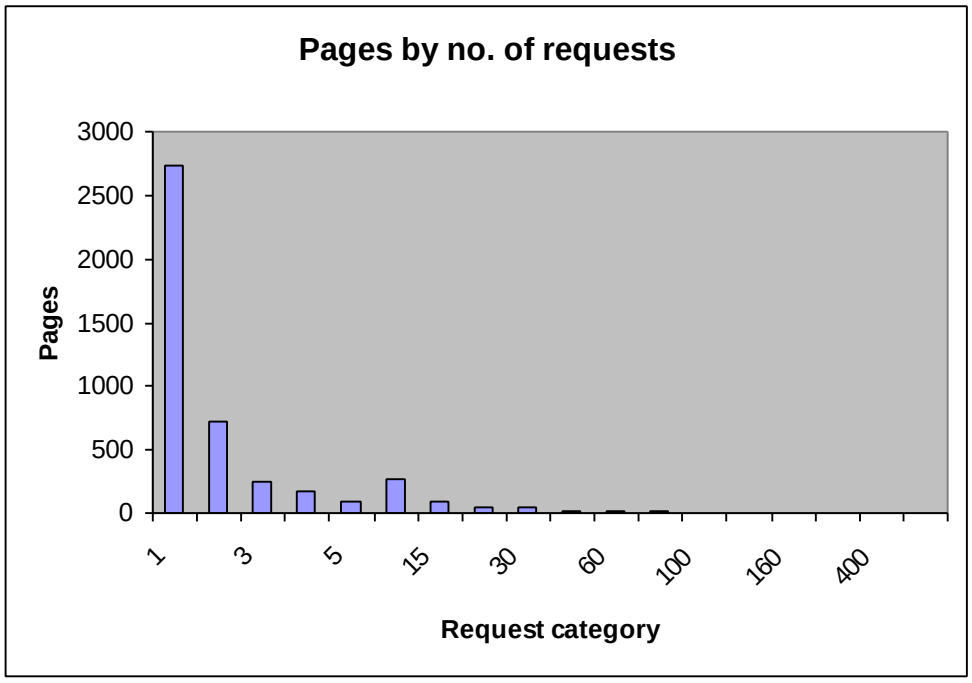


Figure 1: Visitors classified by number of pages they requested.

The popularity of pages can be illustrated by *Figure 2*. It indicates for instance that a majority of the pages, 2727 pages, were only requested once during the day logged. A surprising high fraction of the 4.479 were visited only few times during the day.

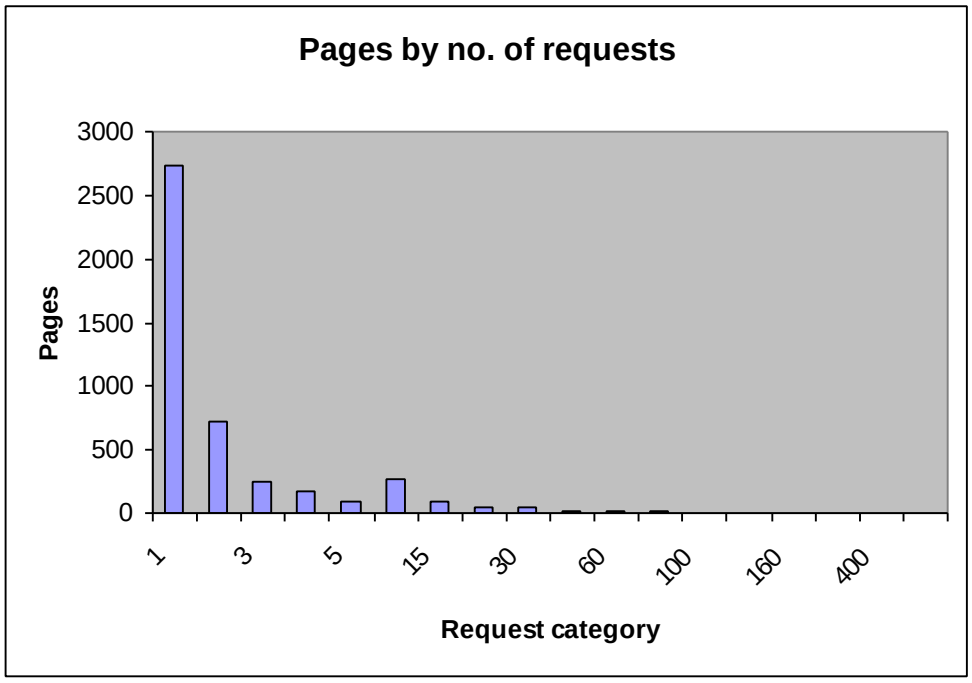


Figure 2: Pages classified by number of requests.

Table 1 exemplifies the length distribution of the visits. All figures are based on the definition that all requests from a client are considered to belong to the same visit, which may of course not be true. There are for instance 24 visits by one crawler each hour for 5 to 12 seconds, which were counted as one visit.

Some visitors were recorded with requests and identifications, which were not properly timed. These records were eliminated in the 'time' analysis. The time distribution tabled included only 1.374 visitors. 326 visitors were recorded with positive visit time 1 second or less. These seem not to represent real visitors and should be excluded. The last 65 visitors were visiting for more than 10 hours. As mentioned above some of these seem to be crawlers which spent a few seconds for catching a set of pages, extracted links from these pages and came back after some time to retrieve a new set of pages.

Visit length	No.of visits
=00:00:00	388
00:00:01 - 00:01:00	326
00:01:01 - 00:02:00	88
00:02:01 - 00:03:00	60
00:03:01 - 00:04:00	44
00:04:01 - 00:05:00	37
00:05:01 - 00:06:00	33
00:06:01 - 00:07:00	22
00:07:01 - 00:08:00	22
00:08:01 - 00:09:00	19
00:09:01 - 00:10:00	15
00:10:01 - 00:20:00	104
00:20:01 - 00:30:00	40
00:30:01 - 01:00:00	38
01:00:01 - 02:00:00	21
02:00:01 - 03:00:00	13
03:00:01 - 04:00:00	7
04:00:01 - 05:00:00	7
05:00:01 - 06:00:00	1
06:00:01 - 07:00:00	5
07:00:01 - 08:00:00	8
08:00:01 - 09:00:00	5
09:00:01 - 10:00:00	6
10:00:01 - 12:00:00	11
12:00:01 - 15:00:00	19
15:00:01 - 18:00:00	8
18:00:01 - 21:00:00	13
More	14

Table 1: Visits by time length

In *Table 2* the requested pages are reduced to the above 1.374 visitor and obvious incorrect request eliminated. The resulting 16.948 'fact' pages are sorted by top level directory indicating a high frequency of requests from '/databaser'.

Number of requested pages by top level directory

Requested URL top level:	# pages
/allman	2
/db	18
/_vti_inf	20
/textdb	22
/webbansvariga	25
/nyheter	31
/arkiv	51
/databeng	94
/info	134
/bibliotek	192
/scbswe	206
/utbildning	573
/snabb	597
/omscb	615
/regioner	649
/publkat	663
/sm	693
/scbeng	714
/arbetsmarknad	769
/scbquery	857
/landmiljo	905
/statinfo	1063
/ekonomi	1285
/index	1312
/press	1670
/befovalfard	1747
/databaser	2041
	16948

Table 2: Requested pages by top level directory

Table 3: Transitions from a top level directory to the next

	To:															
From:	Out	/_vti_inf	/allman	/arbetsmarkna	/arkiv	/befovalfard	/bibliotek	/databaser	/databeng	/db	/ekonomi	/index	/info	/landmiljo	/nyheter	/omscb
/_vti_inf						5					7					
/allman																
/arbetsmarkna	25			498	9	38		22			10	17		4		4
/arkiv	7			2	7	22	3	2				2				5
/befovalfard	141	5		9	1	1260	21	36			37	32	5	7		9
/bibliotek	16				3	1	115	24			5	4	4			6
/databaser	85			16	1	31	1	1578	1	1	44	46	6	5		14
/databeng	9								69		1	3				2
/db	7									1		3				
/ekonomi	64	7		5		17	3	24			800	40	19	4		11
/index	329			77	6	116	8	75	8		96	193	12	9	2	70
/info	11			1		2	5				2	6	62	21	4	9
/landmiljo	19					1	1	2			3	7		759	20	2
/nyheter	1				1								2		1	24
/omscb	46			7	3	9	3	5			6	21	7		4	400
/press	140		1	14	2	35		18	1	1	21	34	1	9		11
/publkat	26	1	1	9		47		10	1		22	10	1	19		13
/regioner	24					5		14	1		16	3	2			2
/scbeng	108	2				4		1	8	2	1	27		1		
/scbquery	31			22	12	43	17	14	2	1	36	19	5	13		10
/scbswe	60			2		5	1	3		3	3	17	2			
/sm	7			3		6	2	16			15	6		3		
/snabb	90			11		11		3	3		38	22				2
/statinfo	52	5		41	1	25	1	147			77	19	3	33		6
/textdb	10											7				
/utbildning	18			40	1	9		11			3	7	1	1		3
/webbansvariga				1	2	3	1				1	14		2		1
Grand Total	1326	20	2	758	49	1695	182	2005	94	9	1244	559	132	890	31	604

/press	/publkat	/regioner	/scbeng	/scbquery	/scbswe	/sm	/snabb	/statinfo	/textdb	/utbildning	/webbansvarig	Grand Total
	1		2					5				20
1									1			2
18	12	2	1	8		5	22	43		31		769
1												51
39	44	3	3	47	4	5	1	33		5		1747
2	2			7		1		2				192
15	5	6	4	19	2	13	1	141	1	5		2041
4			4	1			1					94
			2	2	3							18
37	13	16	1	25	5	14	90	85	1	4		1285
49	20	9	52	115	8	1	40	4	1	11	1	1312
		2		6		1		2				134
7	24	6	1	8	1	6		37		1		905
2												31
38	18	19	1	12	1	1	3	6	1	3	1	615
1175	42	8	41	29	25	7	20	8		6	21	1670
9	402	5	14	32	2	14	7	13		5		663
7		531	1	8	1	3		4		27		649
33	28		395	2	33	1	63		2	3		714
41	15	12	4	487	2	7	4	44		14	2	857
31	2	2	29	2	43		1					206
3	9	2		3	2	602	1	9		4		693
24	5	1	68	5	4	1	300	9				597
16	10	4	1	17	3	7	3	541		51		1063
1			1		2				1			22
8	5	3	1	13	1	2		50		396		573
												25
1561	657	631	626	848	142	691	557	1036	8	566	25	16948

Table 3: Transitions from top-level directories to top-level directories