Evaluation of User Search in a Web-Database

Joan C. Nordbotten
Department of Information Science
University of Bergen, Bergen Norway
e-mail joan@ifi.uib.no,
Web-site: http://www.ifi.uib.no/staff/joan/

Svein Nordbotten
University of Bergen
Bergen Norway
e-mail: svein@nordbotten.com
Web-site: http://nordbotten.com/

Abstract

The number of Web-databases has exploded during the last years. In order to justify the development of new information resources, it is essential to know if the use of existing resources has followed a similar trend. This paper presents an analysis of the use of a national statistical web-database made to support Web site improvement efforts.

The study is based on log data taken during 2 time periods in 1999 and 2000. In this period, the number of registered users increased 5-fold and the number of sessions more than double. During September 2000, active users spent 4,320 hours on the Web-DB and initiated 14,998 topic-sessions giving an average of 7 hours and 25 sessions per user.

Definition of a session has proved difficult since the log data available is based on registered organizations, rather than on tasks or individual persons. Ideally, a session should be defined as a search and retrieval for the information required for a task. We have used a topic-session, defined as the sequence of requests from topic initiation to retrieval of data from this topic, as a task approximation.

1. Introduction

Information made accessible on the Internet has exploded in variety and volume during the last years. However, has the use of the information followed a similar trend? Who are the users? What information do they request? How do they find information? What do they download? Answers to these questions are essential for justifying the development of new information resources on the Net.

This paper focuses on a special, but important information resource, web-databases providing statistical facts about socio-economic conditions. The Web-DB used in this study, the Swedish Statistical Databases (SSD), was established on the Internet by Statistics Sweden (SCB) in January 1997 and is available at www.scb.se. The objective was to make official statistics for Sweden more accessible for users to search and download. For security reasons, all users have to register before getting free admission to SSD. This gives SCB a rich source of data for evaluation of the use of SSD. The purpose of the current 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.

A pilot study indicated that the number of *different* visitors to the SCB web site a day in April 2000 was about *1.500*. These users made about *17.000* requests to the site, of which *2.000* were requests for entry to the database.

The current study is based on user registration data and log data of their use of the Web-database in September 1999 and September 2000. This study is related to other studies we have made focusing on evaluation of the interface to websites [1, 2, 3, 4].

2. Users of the Swedish Statistical Databases, SSD

Users of SSD have been registered since January 1997 and are primarily organizations, such as government agencies, universities, media, business organizations, or libraries. Individuals have been a minor user type. The registration process gives the user an identification number and password, while SCB keeps data on registration date, user classification (business, government, etc.), and geographic location. For this study, the user population has been limited to registered external (to SCB) users at the end of September 2000.

There were *1.555* users registered prior to October 1999, less than *20%* of the *8316* users one year later. During1997-1999, monthly increases were modest. After free access was given in January 2000, the increases were significantly higher. Free access is a likely explanation for the first 3-4

months, while increases in the later months of 2000 are likely to be the impact of the general explosion in Internet use. The trend shown in Figure 1 indicates that the number of users might pass *10.000* before the end of 2000, which it also did.

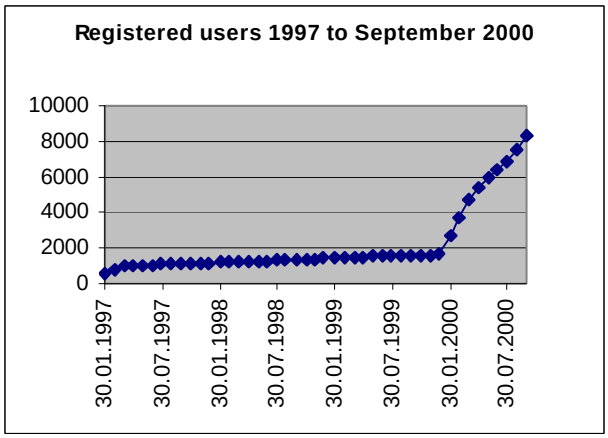

Figure 1: Growth in registered users

2.1 User Categories

The number of registered users, by category, is given in Table 1. Note that about 7% of the total registered users were active in September 2000.

Table 1: Registered and Active Users, Sept. 2000

code	*Category*	*Registered*	*Active*
P	Private citizens	928	106
E	Economic enterprises	4470	219
L	Public libraries	136	33
U	Universities	349	42
S	Schools	148	20
MC	Municipalities/counties	894	65
M	Media	274	21
O	Organizations	257	15
GD	Government departments	28	3
SA	Statistical agencies	24	2
GA	Government agencies	230	26
F	Foreign (to Sweden)	528	48
	Other	50	0
	Sum	8316	600

The distribution of users by category changed significantly during the study period as shown in Figure 2.

In 1999, *Municipalities* were the leading user category accounting for almost *34%* of the registered users. By September 2000, *Economic enterprises* accounted for *54%* of the registered users, while *Private citizens* and *Municipalities* came in 2nd and 3rd place with about *11 %* and 10.5% respectively. Relatively, *Private citizens* increased from about *1%* to *11%*, *Economic enterprises* increased from 26% to 54%, while *Municipalities* went down from *34%* to *10.5%* even though their absolute number increased by *369* to *894*.

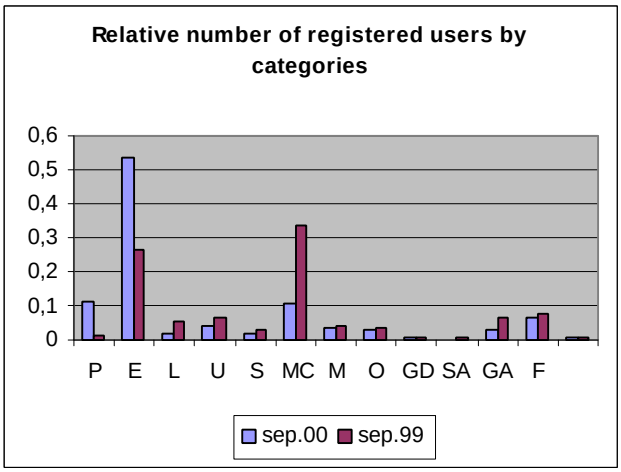

Figure 2: Change in relative numbers of users

The user population has a wide geographic distribution. By September 2000, non-Swedish users accounted for *7%* of all users and represented 48 countries. The trend from the previous year is interesting. While the total number of registered users increased by a factor 5.3 from Sept. 1999 to Sept. 2000, non Swedish users increased by a factor *11,7* from *49* to *571*, indicating a significant growth in interest for statistical information from Sweden among foreign users.

Figure.3 shows the relative user frequencies from countries with more than 5 users. The Nordic countries, USA, UK, and Germany dominate as the location of SSD foreign users. Today, most of the meta-data used for navigation in the Web-DB is still in Swedish, while only *1/3* of the foreign users were from countries with Scandinavian languages. When an English language version of SSD is completed, users from foreign countries are expected to increase.

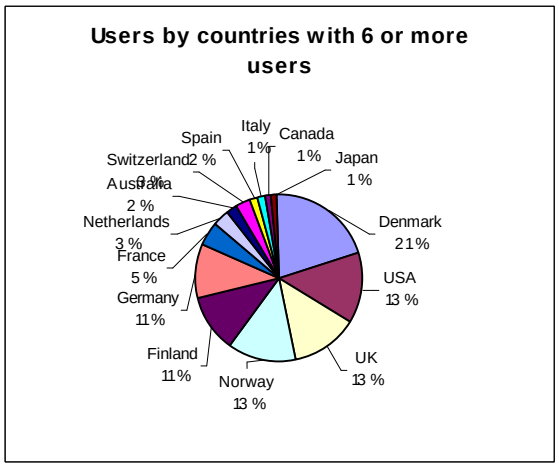

Figure 3: Non-Swedish users

2.2 Secondary users

A user must use the unique identification of his registered organization to gain access to the Web-DB. Thereafter, each selection within the database is recorded as a *request* in the DB log. The log entry records the time the request is received, the organization identification, and the page or service requested.

Six hundred different users (*7%)* accessed SSD on one or more days during September 2000**.** Users in categories *Private citizens, Public libraries, Universities and colleges,* and *Schools* were relatively more active than users in the remaining categories. Each user (organization) in the library and educational categories represents many individuals, or *secondary users,* who may not individually be very active, but collectively make the organizational user active.

The category, *Economic enterprises*, had the largest number of registered (organizational) users and the highest number of the active users. However, the percentage of active users from this category was much lower than other categories. One explanation can be that each economic enterprise has few secondary users.

2.3 User activity

One measure of user activity is a *day-session*, defined as all DB activity by a user (organization) in a particular day [3]. With this measure, user activity can be described in a number of ways, including:

- ?? Usage frequency per time unit,
- ?? Visits made in both September 1999 and September 2000,
- ?? Requests made per visit, and
- ?? Time spent per visit.

2.3.1 Usage frequency. The *600* active users had *1.841* day-sessions during September 2000, giving an average of *3,1* day-sessions. Of these, *110* included only one request, most probably from non-Swedish users or ones without a particular task as a background motive.

There was a significant change in activity from September 1999, when the *243* active users averaged 5.1 day-sessions. The decrease of activity to *3,1* days in 2000, can be the impact of the change in the distribution of users by categories. While the absolute number of users increased from 1999 to 2000, the relative increase may have been greatest in categories with low tendencies to repeated use. Relative increases were particularly marked for the categories *Private citizens, Universities* and *Colleges,* and *Government agencies,* while the large category *Economic enterprises* reduced its sessions per user with almost *50%.*

Figure.4 shows the average number of day-sessions per user in each of the 13 user categories in September 2000. As might be expected, the categories with the highest averages were *Public libraries* and *government agencies.* A probable explanation is that these users are established users of official statistics and ones that also have a number of secondary users with individual tasks for which they need statistical support.

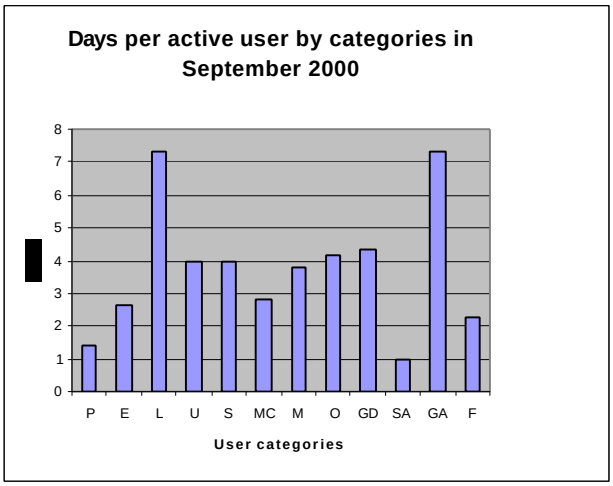

Figure 4: Usage activity per user category

2.3.2 Long-term users. Of the *246* active users in September 1999, *66* or about *25%* were also users in September 2000. Long-term users were particularly strong in *Economic enterprises, Public libraries, Municipalities and counties* and *Government agencies.* While *Public libraries* can be assumed to have many secondary users, users in the 3 other categories have re-occurring tasks, which can explain annual re-visits.

2.3.3 Requests per visit. Another measure of user activity is the extension or length of a session, measured by number of requests per day. The average length of a session was *29* requests, varying from 1 to *761*, with about half between 1-10 requests.

Figure.5 shows the average session length for each user category. The chart shows that users from *Economic enterprise* had the longest average sessions. Though the longest individual session was by a *School* user. A long school session probably consisted of a class of students using their school's user identification. *Public libraries* also rank high, but again the explanation is probably many secondary users.

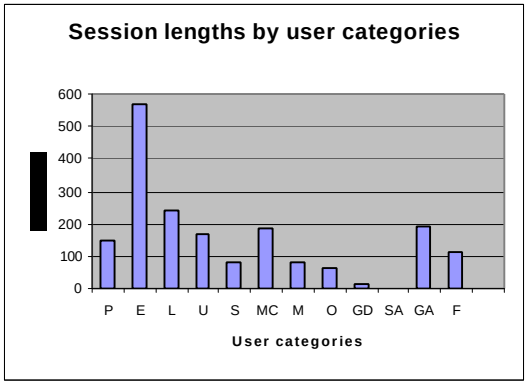

Figure 5: Day-session length in number of requests

2.3.4 Time spent per visit. Time spent on a Web-DB is impossible to measure exactly. Time is recorded at the initiation of a request, and most users leave the site without making use of an exit request. Further, it is impossible to determine whether a user is working with a response, answering a telephone call, or engaged in some other activity between 2 requests. This must be kept in mind when discussing time spent.

In this study, a day-session lasts from the 1st user request to the final request of the day for that user. The *600* active users logged *4.320* hours connection time to SSD in September 2000, giving an average of *7,2* hours each. Day-session time averaged 2,3 hours. The time varied from *0*, for the *178* sessions consisting of a single search request, to the session that lasted for nearly *24* hours. Half of the sessions lasted less than 20 minutes.

Compared with time spent with other types of databases, in which an average time is measured in seconds or minutes [1, 5, 6, 7], a significant number the users spent long sessions with SSD. This most likely reflects a different form for data processing, though a reservation must be made for the uncertainty in the delimitation of secondary user activity.

There were *53.598* requests made in September 2000 giving an average request time of nearly *5* minutes. Also this figure is higher than found in similar studies in connection with other Internet databases. These observations support the idea that users of statistical databases seem to spend more time working with the responses from the database.

Figure.6 shows the relative time spent by user category in September. As expected, time spent correlates with the number of requests per session.

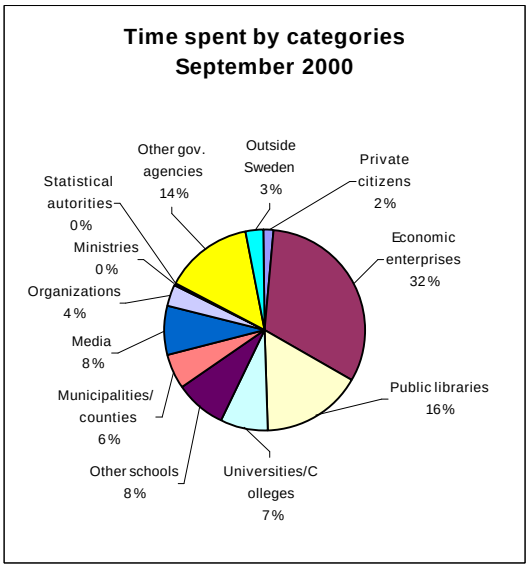

Figure 6: Time spent per user category

3. Statistics requested and retrieved

The objective for establishing a statistical Web-database is to provide users with easy access to relevant statistics that support their needs. It is extremely important to acquire knowledge about the behavior of the users of a Web-DB. Such knowledge may identify needs, indicate the efficiency of the database design, and show possibilities for improving the database.

In SSD, a session starts with either with a key word search request or with a request for meta-data about one of the 19 statistical areas. A hierarchy of *menus* facilitates a user's search for statistics. The highest menu level gives access to 19 statistical *topic-areas*. The second level menus give access to a growing number of tables within each area. There were about 800 statistical tables in September 2000.

3.1 Statistics selected

In September 2000, *Population statistics* were most frequently requested, accounting for 21% of all requests. Thereafter followed *Trade & services* (16%), *Labor* (13%), and *Industrial & commerce* (8%). *Organizations* and *Ministries* requested *Labor statistics* most frequently, while G*overnment agencies* requested *Trade statistics*. All other user categories requested *Population statistics* most frequently.

The least requested statistics, representing only *1%* of the requests, were *Living conditions*, *Environmental protection*, and *Social conditions and services.* Given the publicity these areas receive, the low interest from the statistical users is surprising.

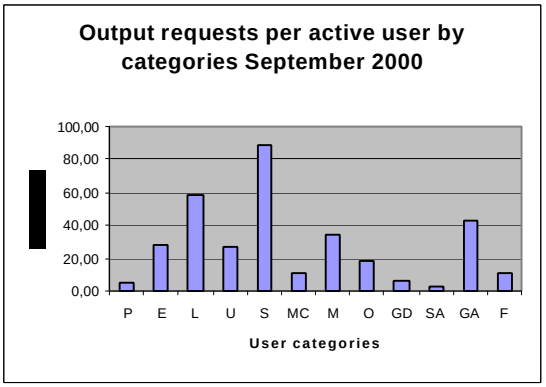

Figure 7: Topic-sessions by user category

3.2 Session analysis

Defining a *session* has proved difficult. Request identification by organization does not directly support a study of task or secondary user activity. This is a particular problem for analysis of usage from library and educational organizations where the number of potential secondary users is high and where there appears to be concurrent sessions.

Ideally, a *task*, defined as a search for task related information through to the retrieval of that information, should define a session. A simplified definition of a task could be one that retrieved statistical data from a single topic area. We have used this simplified concept as an approximation for task analysis.

A *topic-session* is thus defined as the sequence of requests from topic initiation to retrieval of data from this topic [4]. A topic-session typically includes a set of requests for meta-data within a topic of interest and terminates with an *output request* for statistical information. A topic-session is not equivalent to a task session, in as much as a task may well require multiple output requests for data from one or more topic areas.

The number of topic-sessions more than doubled during the study period from 7378 to 14.998, giving an average of 25 topic-sessions per active user organization in September 2000. The average topic-session consisted of 3.6 requests.

Figure.7 shows the distribution of topic-sessions, as defined by output requests, per user category. Users from *Schools* and *Public libraries* made the highest number of requests for statistical output. The chart also shows that the educational system is becoming a major user of official statistics. *Media* and *Government agencies* are two other heavy users.

3.2.1 Data retrieval. An output request returns statistical information in the form of a display on the user screen or a downloaded file. Display of statistics accounts for *2/3* of all output requests. There were no significant differences between user categories in download frequencies.

While the total number of requests, for both metadata and statistical output increased with a factor of 3.7 from September 1999 to September 2000, output requests doubled for the same period. 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 from *Private citizens.*

There was a fall, from *10.628.037* to *6.126.903* rows of data downloaded in September 1999 and September 2000 respectively, while there was a small increase in the volume of statistics displayed on the users' screens. Explanations for the reduction in the volume of output can be many, e.g. the users may have become more focused and knew what they wanted.

The average number of rows retrieved per day per user in September 2000 was *3.326*, which is a surprisingly large row number. Figure.8 shows how the downloaded statistics, measured in rows, are distributed by user category. Users from *Outside Sweden* had a particularly high volume of output per session. One explanation can be that these are foreign trade representatives or trade attachés in foreign diplomatic agencies, with a regular task to report statistics from Sweden.

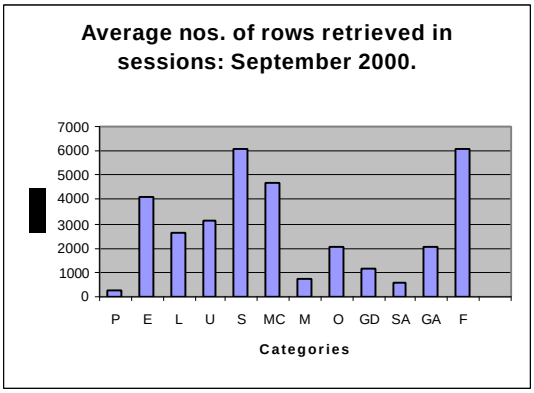

Figure 8: Statistical data requested by user category

4. Summary and Implications

We initially asked the following questions about Web-DB use.

- ?? Has the *use* of Web-DB information increased along with the increase in Internet databases?
- ?? Who are the users?
- ?? What information do they request?
- ?? How do they find information?
- ?? What do they download for local processing?

Our study has been based on user registration and log data from the Swedish Statistical Databases (SSD), established on the Internet by Statistics Sweden (SCB), and available at www.scb.se. Our observations are summarized below.

4.1 User interest

The number of users has increased by a factor of *5* from September 1999 to September 2000, due partly to establishing free access, as well as to the increasing use of the Internet.

Topic-sessions, measured by requests for statistical data, more than doubled during the same period.

The average number of day-sessions per user decreased from *5,1* in September 1999 to *3,1* in September 2000. Possibly, the cancellation of the annual fee has attracted users without an economic motivation to exploit the database. This trend can be expected to continue as the number of non-economically motivated users increases.

4.2 User identification

By the end of September 2000, *Economic enterprises* had became the dominant user category, representing more than *50%* of all registered users and reducing *government* users to 2nd place. Enterprise users appear to have found an easily accessible source of statistical information for business decisions. Their number can be expected to continue to grow.

Private citizens now represent 11% of the user population, up from 1% in 1999. This number is expected to grow.

Foreign (non-Swedish) users have increased by a factor of 12, to 7% of the user population, during the study period, indicating a growing interest in national statistics.

4.3 User Activity

About *7%* of the registers users accessed SSD in September 2000, up from *4%* in September 1999.

Public libraries and *Government agencies* were very active users in September 2000, logging over 7 topic-sessions per user. Users in these categories have a potentially large number of secondary users.

Users from *Universities and colleges* increased their activities relatively more than any other user category. Given free access and a high number of secondary users, this trend can be expected to continue.

Economic enterprises represented about 1/3 of the active users, but ranked 5th in number of topic-sessions and 9th in average day-sessions. This may be because these organizations have few secondary users and a more focused need for statistical data.

27% of the active users in September 1999 were active in September 2000. This is a surprisingly high fraction that probably represents long-term users.

4.4 Information requested

Population statistics were most frequently requested by 10 of the 13 user categories, and accounted for 21% of all requests. Thereafter followed *Trade & services* (16%), *Labor* (13%), and *Industrial & commerce* (8%).

4.5 Information location

In September 2000 there were a total *53.598* requests to the database of which *14.998* were requests for statistical output. The remaining 72% of the requests were for meta-information to identify the statistical topics.

The average topic-session contained 2.6 requests for metadata for each request for statistical output. Request time averaged 5 minutes indicating a relatively long on-line processing time for information location and retrieval.

The average length of a day-session was *29* requests including about *10* output requests. Session lengths varied from *1* to 761 requests. Session time averaged 2.3 hours.

4.6 Information retrieval

Retrieval requests in September 2000 averaged *3.326* rows per output. About 2/3 of the data were downloaded for local processing.

School users had twice the overall average in output requests.

Though output requests doubled from September 1999 to September 2000, the absolute volume decreased by 42%. A probable explanation is the increase in smaller tasks from *educational organizations* and *libraries.*

5. Conclusion

Our study shows that, at least for statistical Web-DBs, the increase in information available on the Internet is followed by an increase in usage. We have observed a 5-fold increase in the number of users and a doubling of output requests during a 1-year period. New users groups have appeared and there is no indication that this activity will stop.

Acknowledgments

This paper is based on a study from 2001 commissioned by Statistics Sweden (SCB) [3.]. The authors are grateful to SCB for permission to present this paper.

References

[1] Nordbotten, J. and Nordbotten, S. (1999). *Search Patterns in Hypertext Exhibits.* HICSS-32. Proceedings of The Thirty-Second Annual Hawaii International Conference on System Sciences. Maui, HI. January 4-7 1999. IEEE. ISBN 0-7695-0001-3.

[2] Nordbotten, S. and Nordbotten, J. (2001a): *Perception of Statistical Presentations Investigated by Means of Internet Experiments.* Hicss-34. Proceedings of the 34th Hawaii International Conference on System Sciences. January 3-6 2001. IEEE. ISBN 0-7695-0981-9.

[3] Nordbotten, S. and Nordbotten, J. (2001b): *A Study of the SSD Web Logs for September 1999 and September 2000* - A Report to Statistics Sweden, Stockholm.

[4] Nordbotten, J. and Nordbotten, S. (2001c): *Student Usage of a Statistical Web-DB.* Proceedings of the HCI'01, New Orleans. August 5-10, 2001.

[5] Shneiderman, B., et.al. (1989). Evaluating Three Museum Installations of a Hypertext System. *Journal of the American.Society for Information Science,* **40**(3), 172-182.

[6] Shneiderman, B. (1998). Designing the User Interface - Strategies for effective Human-Computer Interaction, 3rd ed. Addison-Wesley.

[7] Yamada, S., et.al. (1995). Development and evaluation of hypermedia for museum education: validation of metrics. *ACM Trans. of Computer-Human Interaction,* **2**(4), 284-307.