



ALIA 2006 Biennial Conference



Australian Library and  
Information Association

## Refereed Paper

### Markus Hennies

Associate Professor, Stuttgart Media University, Faculty of Information and Communication

#### Contact details

Postal: Stuttgart Media University, Wolframstr. 32, 70191 Stuttgart, Germany

Email: [hennies@hdm-stuttgart.de](mailto:hennies@hdm-stuttgart.de)

#### Biography

Professor Markus Hennies graduated in physics (1989) and management (1999). From 1993 until 2004 he was head of Information Systems Management at Freiburg University Library. Since 2005 he has held the position of Associate Professor at Stuttgart Media University, Faculty of Information and Communication. He specialises in library and information management studies and is conducting research on information retrieval systems. In 2004 he was awarded the Graduate Teaching Award of the State of Baden-Württemberg, Germany. He is a member of various state and national library committees/research groups such as the Working Committee on Electronic Interlibrary Loan, State of Baden-Württemberg, Germany and the Research Group Web and Databases of the German Informatics Association (GI).

### Juliane Dressler

Heidelberg University Library

#### Contact details

Postal: Heidelberg University Library, Ploeck 107-109, 69117 Heidelberg, Germany

Email: [dressler@ub.uni-heidelberg.de](mailto:dressler@ub.uni-heidelberg.de)

#### Biography

Juliane Dressler, studied library and media management at the Stuttgart Media University from October 2001 till February 2005. She spent a practical term at the Stockholm Public Library and worked as a trainee at several libraries in Germany and

at a company specialised in library software. In her diploma thesis she analysed the transaction logs of the online catalogue of the Freiburg University Library. Since March 2005 she has been working at the Heidelberg University Library, especially at the department library for economics and at the information service. She is particularly interested in customer orientation and in the application of information technology in libraries.

## **Clients information seeking behaviour: An OPAC transaction log analysis**

### **Abstract**

A study about information seeking behaviour of students and staff using OPACs is presented. Based on about 6 million requests sent to the OPAC of the Freiburg University Library system from March to July 2004, this study was created by applying transaction log analysis (TLA). Initially the data acquisition procedure is described. The original web server log was expanded to provide additional fields. Then the log entries were sorted into request types identified by the referrer field. The OPAC usage in the course of an average day was aggregated for a number of user groups.

In a next step the query structure was explored (e.g. the number of query terms and fields used as well as the usage of operators and truncations). By comparison with a similar TLA effects of the interface layout on the queries could be identified. In the last part of the study another approach based on heuristic session analysis dealt with subsequent queries within a session and the clients navigation within the result sets. The number of queries per session, the number of full titles displayed and the number of short title pages viewed was extracted from the log file.

### **Introduction**

The rise of the online catalogue about 30 years ago made it possible for the first time to objectively examine the information seeking behaviour of a great number of catalogue users. Until then, researchers had to rely on questionnaires, direct observation or they had to work with focus groups. The method applied to examine the information seeking behaviour directly in the OPACs is called 'transaction log analysis' or 'TLA'. Since TLA seeks to analyse uninfluenced search behaviour it can be very useful for designing and evaluating catalogue user interfaces. The advantages of this method lie in the fact, that the acquisition of all the data needed can be done rather easily and complete, and that these log data can be used for answering different kinds of questions. The limitations of this method lie in the lack of demographic information about the catalogue users and the lack of feedback from the users regarding the qualitative aspects of the use (such as reasons for the search and satisfaction with the search results) on the one hand and on the other hand in the fact that analysis in TLA is rather time-consuming (for an overview of the use of TLA methodology in libraries see Covey, 2002, pp. 33-44)

This article analyses the web server log of the OPAC of the Freiburg University Library system over a period of twenty two weeks from March until July 2004 which comprised 6,098,620 data sets.

The Freiburg catalogue can be accessed via <http://www.ub.uni-freiburg.de/olix> and contains two million records of holdings from different institutions such as Freiburg University (25,000 students, 1,600,000 records) and several other academic libraries of local colleges of TAFE (5,000 students).

### **OPAC user interface**

The Freiburg OPAC can be accessed without prior registration. The catalogue search screen provides three search modes: Standard search, expert search and index search. In standard search the user can fill in up to three fields which are pre-defined to search for title keywords, author and subject heading. In this search mode the fields are combined with the default operator 'AND'. Via pull-down menu the pre-defined search fields can be changed to different aspects and the operators can be changed to 'NOT' or 'AND NOT'. In expert search mode more complex searches such as nested searches can be entered directly in command language. The third mode is a direct search in the index of all the entries in the catalogue.

A successful search in the OPAC of the Freiburg University Library system results either in an exact hit which is immediately shown in full title display or in multiple hits displayed as a short title list of twenty titles per page. Each of the short titles shown in the list can be expanded to full title display.



Figure 1: OPAC search screen providing three search modes

**Data acquisition**

The Freiburg OPAC technology is based on a multi-tier software architecture made up of the presentation tier, the application tier and the data tier. The presentation tier is realized by a 'CGI script' (Common Gateway Interface script) running on the

web server. This script translates the original query into command language and sends it to the application tier which in our case is the catalogue server. The response of the catalogue server is translated by the CGI script into a HTML page which is finally sent back to the client's browser.

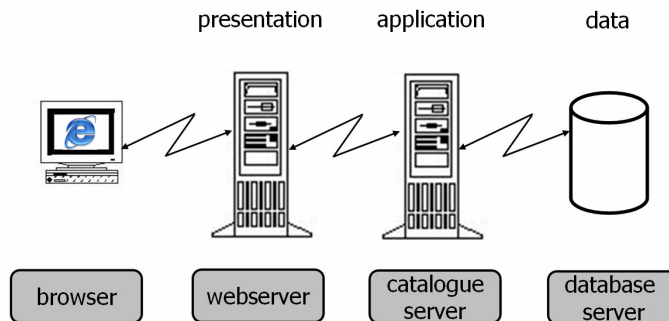


Figure 2: Multi-tier architecture

Transaction logs can be generated on either the presentation tier (web server) or the application tier (catalogue server). By default web servers like the Apache httpd create log files of the web pages requested, containing client's IP address, a time stamp of the request, the path to the object requested, a status code of the response and the

number of transferred bytes. However, web servers neither log the parameters of the query (if passed by method POST) nor the number of hits achieved. Most TLA studies therefore analyze logs generated by the catalogue server. But these logs lack information on the origin and the details of the request like for example the search mode. Furthermore

they cannot provide information on whether the log entry was created by a query entered directly into the search screen or by navigation within a result set. For our TLA we therefore enhanced the information acquired from the web server by information gathered from the catalogue server. This was realised by expanding the CGI script with an additional logging routine. Our customised log file thus contains unique ID, time stamp, client IP address, referrer page, request parameters and the number of hits.

**General analysis**

In a first step we sorted the logged requests by the content of the referrer field. This field contains the URL of the webpage the request originated from. We found that 2,329,755 (38.2%) requests originated from direct entries by users into the search screen, 3,368,202 (55.2%) from browsing in result sets, 244,964 (4.0%) from book shelf number requests, and the rest resulted from other links. Since users of the catalogue of the Freiburg University

Library system do not have to register prior to searching the OPAC we tried to identify target groups by their IP addresses. We were able to identify different user communities of Freiburg University and the local colleges and took a closer look at the distribution of their requests in the course of a day. These communities accessed the OPAC from following major sites: 2,346,108 (38.6%) from public access PCs within the University Library, further 1,147,629 (18.8%) from within the University network, 873,189 (14.3%) from other Freiburg colleges of TAFE and 1,721,694 (28.2%) from elsewhere.

Analysing the average daily OPAC usage from the library's public access PCs it becomes clear that the library opening hours (8 AM to 10 PM) are reflected by the average daily OPAC usage. The highest usage level overall is reached between 11 AM and 4 PM. Interestingly to observe is the slight dent around lunch break from all sites except from the public access PCs within the University Library (figure 3).

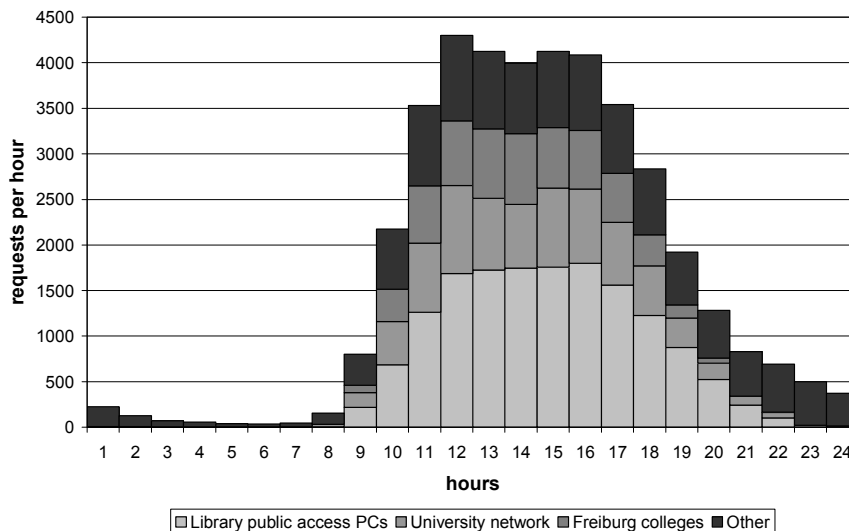


Figure 3: Average daily OPAC usage

### **Query analysis**

Our query analysis focused on the 38.2% (2,329,755) of requests originating from direct entries by users into the search screen. Of these requests 96.1% (2,239,179) resulted from standard search mode, 2.3% (54,429) from index search mode and another 1.6% (36,084) from expert search. In the following we are presenting our analysis of the queries entered in standard search mode only.

In standard search mode of the Freiburg University Library system OPAC up to three search fields can be filled in. 80.2% (1,795,633) of the queries entered were single field queries, 19.1% (427,653) were two field queries and only 0.7% (15,714) were three field queries. 99.3% of the combined searches with two or three fields used the pre-set ,AND' operator. 0.6% used the ,OR' operator and only 0.1% used the operator ,AND NOT'.

In standard search mode the search fields are pre-set to title keyword, author and subject heading. The order in which these fields are arranged on the search screen is reflected in the frequency the fields are used: The title keyword field was used in 53.3% (1,193,996) of the queries, the author field in 39.5% (885,271) and the subject heading field in 16.9% (378,385) of the queries. Additional fields, such as book shelf number, free text, publishing year, classification and publisher, which could be selected via pull-down menu, were only marginally used (in less than 5% of the queries).

Further analysis of the queries in standard search mode showed, that in the vast majority of queries the default

values for search fields and operators remained unchanged (85.3%). This confirms earlier findings by other studies (e.g. Jones, Cunningham, McNab, Boddie, 2000, pp. 155-156). As Yee and Layne (1998, p.1) summarize: "Catalog use research reveals that users rely heavily on defaults."

Going a step further, we took a closer look at the search field entries. The title keyword field as well as the free text field allow for multiple entries. However, 42% of the queries using these fields were single word entries. Two words were entered in 26.4%, three words in 16.6%, four words in 7.9% and more than four words were only entered in the remaining 6.8% of the queries. The record number of 159 words was entered in title field by one catalogue user, obviously copying a complete abstract into the search field.

The Freiburg OPAC allows truncation. Only in 2.9% of the queries this option was used at all. However, instead of using the correct truncation sign ,?' every tenth truncated query used an invalid truncation sign such as ,\*', ,#', ,\$', ,%' . Not surprisingly, user groups using truncation above average show invalid truncation less often.

### **Comparison with similar study**

A similar TLA study was conducted by Remus (2002) for the OPAC of the University Library at the European University Viadrina, Frankfurt/Oder (<http://ubopac.euv-frankfurt-o.de:8080/webOPACClient/start.do>). He logged OPAC usage data for a period of eleven weeks in 2002. This study was chosen for comparison due to the fact that the OPAC user interfaces of Frankfurt/Oder and Freiburg are very

similar. Both provide the three search modes standard search (with three pre-set search fields), expert search and index search. In standard search mode the only major differences between the two OPACs occur in the order of the search fields (author and title keyword are presented on the search screen in reversed order) and the way the operators are presented. On the Freiburg OPAC search screen the operators can be selected via pull-down menu, on the Frankfurt/Oder OPAC search screen all three operators are displayed directly on the screen and selections can be made via radio buttons.

We compared Remus' findings regarding the use of search fields and operators to our results (tables 1, 2). We found the reversed order of the pre-set search fields reflected in the order of search fields used by the clients. While in Freiburg the title keyword field was used for 53.35% of the queries the same field was used in Frankfurt/Oder for only 43,1% of the queries. The author field was used for only 39.5% of the queries in Freiburg while the author field in Frankfurt/Oder was used for 57% of the queries. We found the usage of the other search fields to be almost identical.

Field	Freiburg	Frankfurt/Oder
Title keyword	<b>53.3%</b>	43,1%
Author	39.5%	<b>57.1%</b>
Subject heading	16.9%	17.4%
Book shelve number	2.4%	1.4%
Publishing year	1.2%	2.3%
Classification	1.0%	0.7%

Table 1: Usage of search fields in standard search mode

In standard search mode the more obvious presentation of the operators in Frankfurt/Oder led to a higher usage of the non-default operators ,OR' and

,AND NOT' compared to Freiburg. Nevertheless, in expert mode the operator ,OR' was used more often in Freiburg than in Frankfurt/Oder.

Standard search mode	Freiburg	Frankfurt/Oder
AND	<b>99.3%</b>	90.9%
OR	0.6%	<b>8.5%</b>
AND NOT	0.1%	0.6%
Expert search mode	Freiburg	Frankfurt/Oder
AND	83.0%	<b>99.2%</b>
OR	<b>15.1%</b>	0.4%
AND NOT	1.9%	0.4%

Table 2: Usage of operators in combined searches

### Session analysis

The OPAC of the Freiburg University Library system does not provide session control. In order to be able to analyse a session, we used a heuristic approach. We identified log entries belonging to the same session by examining the IP address and the timestamp of the request. For this approach we had to exclude log entries from identified proxy servers (since proxy servers show different clients as originating from a single address). Unfortunately this procedure affected the largest user group. Requests originating from the public access PCs in the University Library of Freiburg are routed via a proxy server and thus could not be considered for session analysis. Furthermore we excluded entries from special user groups such as library staff and training courses (since they tend to show non-average searching behaviour). In the end the session analysis processed 3,085,497 log entries. Of those 1,049,412 could be identified as queries and 1,535,504 as navigation in

the result sets of a specific query. The remaining navigation entries did not match a query (e.g. due to tabbed browsing or the use of the back button in the browser) and were excluded from analysis.

### Queries per session

Two subsequent queries from the same IP address were assumed to belong to the same session as long as the time elapsed between them was less than a specific inactivity interval. Reasonable values for that interval are ten to thirty minutes, we choose the upper limit of thirty minutes, which was also used in other TLA studies like the one by Mahoui and Cunningham (2001, p. 16). The average session contained 5.5 queries (median 3 queries). 31.4% of the sessions were one query sessions, 17.5% were two query sessions. More than ten queries occurred in 12.7% of the sessions (figure 4). The average time period between two subsequent queries in a session turned out to be 126 seconds.

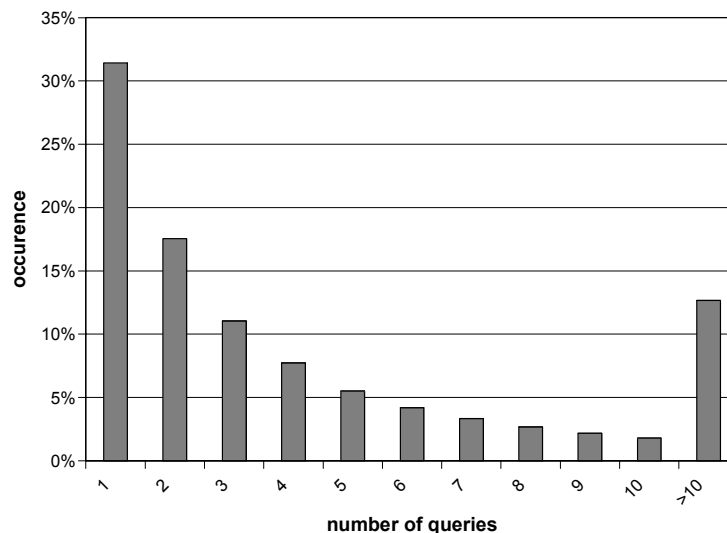


Figure 4: Number of queries per session

**Hits per query**

Zero hit searches were dominating with 27.0% of the queries, whereas 13.7% of the queries resulted in a single hit. Another 26.2% of the searches yielded between two and ten hits. The average

number of hits per query was found to be one hundred and fifteen, but with the median lying at three hits per query. The record number of hits was 1,517,867 as the result of a query with the language as only search aspect.

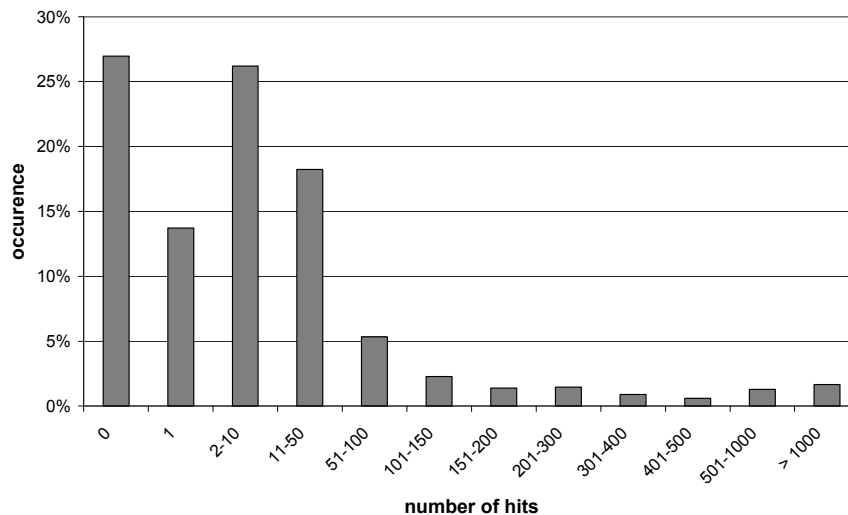


Figure 5: number of hits per query

**Navigation within results**

In the case of queries with more than one hit the resulting titles are ranked by publication year. Journal titles and serial titles are automatically sorted to the end of the result list. The OPAC of the Freiburg University Library system only allows linear browsing in the results.

Looking at the successful queries resulting in at least one hit we determined three characteristic values: the number of full title displays, the rank of the endmost displayed full title and the number of short title pages displayed.

Surprisingly, in 34.5% of the successful queries no full title was displayed. 50.5% of the successful queries led to the display of exactly one full title (18.7% of those due to a single hit query) (see figure 6). On the average 1.1 full titles were displayed per query.

Another interesting finding was the rank of the endmost full record displayed. Figure 7 shows that only in 11.1% of the successful queries a full record beyond the first ten hits is displayed.

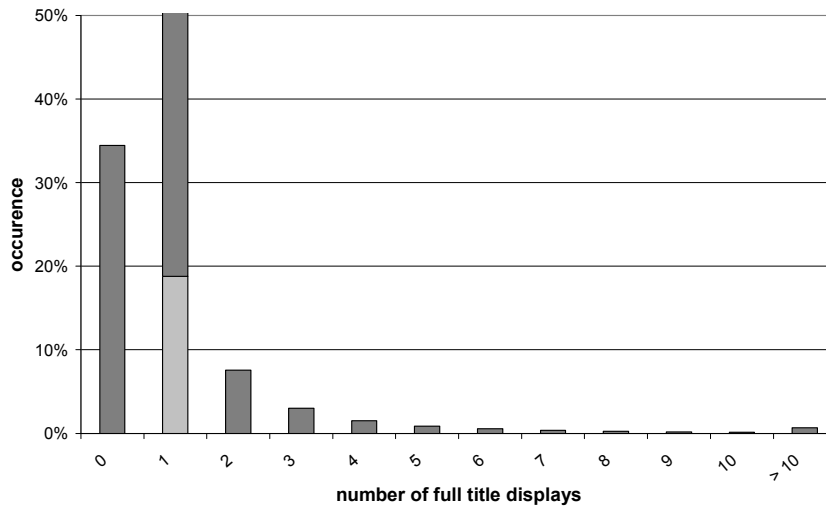


Figure 6: number of full title displays

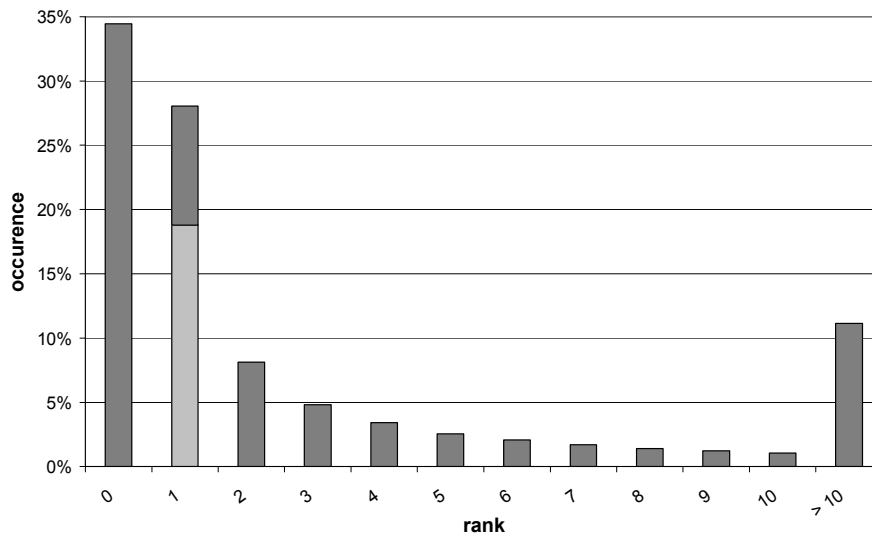


Figure 7: Rank of endmost full record displayed

Queries with at least two hits result in the display of a short title list of twenty titles per page. Figure 8 shows depending on the number of hits how many pages of this short title list were

browsed on the average. Only if the number of hits exceeded two hundred, two pages of the short title list were displayed on the average.

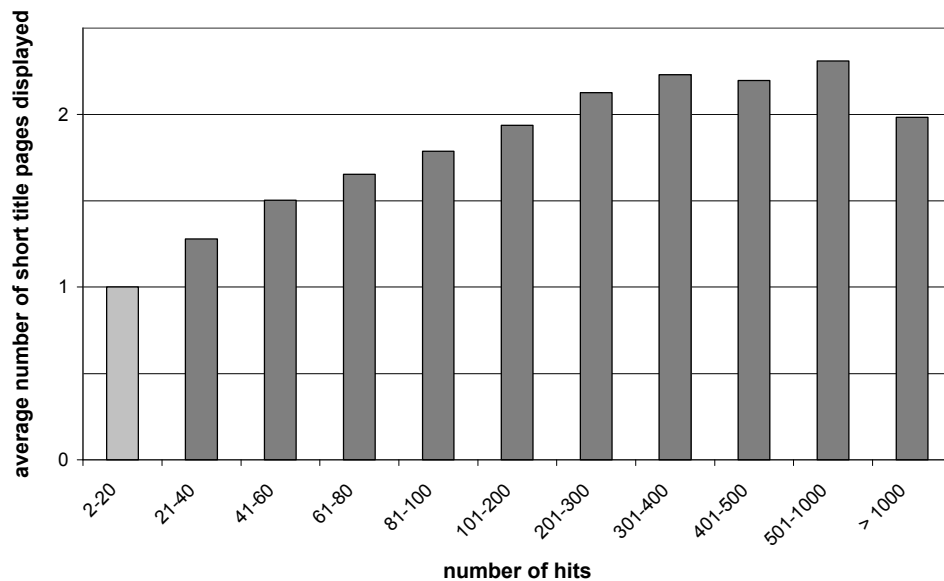


Figure 8: pages of the short title list displayed

### Summary

Different aspects of OPAC usage were studied using TLA methodology. 39% of the requests to the OPAC of the Freiburg University Library system originated from public access PCs within the library. Off-campus use adds up to 28%. The length of an average search session was 5.5 queries, every third session was a one query session.

Regarding the search interface our analysis confirms findings by other authors. Catalogue users tend not to change defaults. In Freiburg the use of the standard search mode dominates in 96% of the queries. Default search fields and operators are kept unchanged in 85% of standard mode searches. Direct comparison with another TLA study showed that usage of non-default operators can be increased by presenting the additional operators as radio buttons instead of pull-down lists.

If truncation was used at all in the Freiburg OPAC every tenth truncation effort used an invalid sign and therefore resulted in zero hits. The automatic translation of truncation signs known to users from other common search interfaces into the valid truncation sign of the OPAC could help to decrease the number of unsuccessful searches.

One in four queries resulted in zero hits. Looking at the browsing behaviour of successful searches with at least one hit, we found that every third search users did not look at any full title display. In only one out of nine successful searches full titles ranked eleven or higher on the short title list were displayed.

Interestingly enough, we found parallels with TLA studies of Internet search engines. The part of one query sessions in search engine usage is slightly higher than in our study and thus the average

session length shorter (Jansen & Spink 2006, p. 255). The number of query terms entered cannot be compared due to the fact that Internet search engines do not allow for structured queries as

library catalogues do. Regarding browsing behaviour the findings match again. In most cases only hits from the first result page are looked at (Spink & Jansen, 2004, chap. 6).

## References

- Covey, D. T. (2002): *Usage and Usability Assessment: Library Practices and Concerns*. Washington: Digital Library Federation, Council on Library and Information Resources. Retrieved June 30, 2006, from <http://www.clir.org/pubs/reports/pub105/pub105.pdf>
- Jansen, B. J., & Spink, A. (2006): How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42, 248-263. Retrieved June 30, 2006, from [http://ist.psu.edu/faculty\\_pages/jjansen/academic/pubs/jansen\\_searching\\_the\\_web.pdf](http://ist.psu.edu/faculty_pages/jjansen/academic/pubs/jansen_searching_the_web.pdf)
- Jones, S., & Cunningham, S. J., & McNab, R., & Boddie, S. (2000): A transaction log analysis of a digital library. *International Journal of Digital Libraries*, 3, 152-169.
- Mahoui, M., & Cunningham, S. J. (2001): Search Behavior in a Research-Oriented Digital Library. In P. Constantopoulos & I. T. Sølvsberg (Eds.), *ECDL 2001*, LNCS 2163 (pp. 13-24). Heidelberg: Springer
- Remus, I. (2002): *Benutzerverhalten in Online-Systemen. Eine Transaction Log Analysis an der Universitätsbibliothek der Europa-Universität Viadrina in Frankfurt (Oder)*. Potsdam: University of Applied Sciences, diploma thesis. Retrieved June 30, 2006, from <http://people.freenet.de/Remus/TLA.htm>
- Spink, A, Jansen, B. J. (2004): *Web Search: Public Searching of the Web*. Dordrecht: Kluwer
- Yee, M., & Layne, S. S. (1998): *Improving Online Public Access Catalogs*. Chicago: American Library Association