

A story told by Nexus transaction logs: what to make of it

Henda van der Berg
Programme Leader: Nexus Database System
Tel: 27 12 481-4016
Fax: 27 12 349-1179
E-mail: henda@nrf.ac.za

Ina Fourie
Associate Professor
Department of Information Science, University of Pretoria, Lynnwood Road, Pretoria,
0002
Tel: 27 12 420-5216
Fax: 27 12 362-5181
E-mail: fouriei@postino.up.ac.za

1 Introduction

Over many years transaction log analysis (also called log analysis, log file analysis, or log tracking, and more lately web logging, web log file analysis, and web tracking) has been used to collect information on how information systems such as online library catalogues (OPACs), online and CD-ROM databases, and web-based products and systems are used. The first reports on such transaction logs date back to the 1980s studies on OPACs (Nicholas et al 1999:265).

Different types of analysis can be identified, for example the evaluation of system performance, user behaviour, a combination of system performance and user behaviour, and the use made of operational information retrieval systems by end-users (Peters 1993). Each type of analysis offers different data that is of important in improving information retrieval systems such as databases and websites. An increased interest in users' behaviour when searching web-based systems is clear from the increase in the number of research reports (Fourie 2002; Jansen & Pooch 2001). The purpose of such research studies is normally to improve the design of user interfaces, search engines, navigational features, online help and intelligent agents, information architecture, content description and metadata, the teaching of information skills and the refinement of information retrieval (IR) research methodology (Sandore 1993; Fourie 2002).

The World Wide Web (WWW) has made it possible for more people to gain access to information, for example through search engines, directories, websites, intranets, and portals. Traditional, structured text databases, such as the databases offered by Dialog, NISC or SilverPlatter, are now also available through the web, which means that more people have access to these databases. The same is true of the database which is the focus of this study, namely the Current and Completed Research Project Database (CCRPD) within the Nexus Database System.

Given the general increase of access to web-based information retrieval systems, as well as the interest from a wider population of users, Fourie (2002) believes that it is essential to increase our research efforts on web information seeking/searching behaviour: "It is clear that substantial web information-seeking/searching studies are necessary to refine our knowledge of web information spaces, their design and maintenance and training-related issues." One method of studying web information seeking/searching behaviour is to monitor transaction logs. Nicholas et al (1999:263)

declare: “With the web being such a universally popular medium, accounting forever more of people’s information seeking behaviour, and with every move a person makes on the web being routinely monitored, web logs offer a treasure trove of data. This data is breathtaking in its sheer volume, detail and potential.”

1.1 Scope of paper

The purpose of this paper will therefore be to use an analysis of transaction logs for the CCRPD for the period 1 to 28 February 2002 to show how such data can be used to improve user interfaces and user training amongst other things. The paper will also comment on transaction logs, their use, value and pitfalls. A brief overview will be given of Nexus and the CCRPD, the actual study and collection of data, the interpretation of data and the preliminary recommendations based on the research findings. The paper will conclude with suggestions for further studies and the use of a combination of methods to collect research data on information seeking/searching behaviour.

2 Nexus and its potential value for the South African research community

This paper focuses on the Current and Completed Research Project Database (CCRPD) within the Nexus Database System that is maintained by the National Research Foundation (NRF) in South Africa. It was developed to provide a service to the research community in South Africa, which has special needs because of a historical context of disadvantaged tertiary institutions and/or researchers.

The Nexus Database System consists of various databases, namely the CCRPD, Research Networking Database that contains fields of specialization and contact details of researchers in South Africa, Research Organisations Database, Professional Associations Database, Periodical Submission Requirements Database, Talk Conference Database for forthcoming conferences and the Women-in-Research Database. The set of databases facilitates research capacity development and is therefore a value addition to the whole chain within the research process. It plays a crucial role in research networking and knowledge transfer in the system of innovation in South Africa.

Nexus is available on the Internet at <http://www.nrf.ac.za/nexus>. All the databases have free access, except the CCRPD, which requires a user id and password. The clientele of Nexus includes educational institutions, government departments, R&D planners, science councils, parastatals, research institutions, and media and information intermediaries.

The following benefits and value addition through the Nexus database system can be listed:

- Avoid duplication of research activities
- Establish priorities for research activities
- Act as a source for scientific publishing
- Facilitate exploitation of research results/findings
- Analyse research trends
- Identify potential partners for joint research projects
- Locate persons and organizations with desired skills
- Identify gaps in research fields

It is therefore essential that the database system should allow for user-friendly and effective information searching (e.g. high recall and precision).

2.1 CCRPD

The CCRPD provides bibliographic references of research projects from 1919 onwards, including master's and doctoral studies in all the science domains. It is mainly used to avoid duplication in research activities, to identify studies for re-evaluation of theories and practices, and to keep researchers abreast of new developments, gaps and trends in their field of study. It is also used by R&D planners, media and policy makers. It is therefore of the utmost importance that all relevant records should be retrieved with high relevancy (i.e. high recall and precision ratios are essential).

Until March 1998 the CCRPD was only available upon payment of a subscription fee. Since April 1998 the tertiary educational sector in South Africa has had free access to the password-protected CCRPD through their library systems. It was mainly librarians who executed searches on the database, although recently a limited number of end-users (academics and students) have also been able to obtain access via the intranets of their institutions. Students can also access the database through a special password, and up till now 7794 incidents of access have been monitored.

When designing the interfaces for the Nexus Database System, it was realized that the users may have certain problems because of the levels of computer, Internet and information literacy required by students and researchers. The historical context of disadvantaged tertiary institutions made it clear that Nexus needed to design its interfaces for an extremely heterogeneous group of users. Thus the levels of knowledge users bring to the interfaces were critical factors that can be divided into two categories: technical knowledge and conceptual knowledge. These were taken into consideration when designing the first interfaces and developing training material. The information searcher's level of experience in information retrieval was crucial and users were therefore categorized as novice and or experienced users. Initially Easy and Advanced Keyword Search Interfaces were offered. Both are menu-driven interfaces. End-users are normally considered as novices, while intermediaries are regarded as experienced users.

It is, however, important to monitor the behaviour of users so as to assess the utility of the databases and user-interfaces and to ensure a higher degree of user-friendliness and effectivity of information retrieval in this regard. The rationale for the current study is to check on the effectivity of the changes made after the 1999 study. It was also prompted by the changes in the target group searching the database, and their presumably different levels of search and subject experience. A study was therefore conducted in 1999, which is followed up by the current study.

2.2 Building on previous studies

In 1999 Van den Berg (2000) reported on a study at the CRIS2000 conference in Helsinki. The analysis of the transaction log files was made on the Easy and Advanced Keyword Search Interfaces.

The 1999 study analysed the transaction log files with zero results in order to identify problems experienced with the search interfaces or conceptualization of search strategies. It distinguished between true (there are no records on the search strategy) and false results (the result is zero, because of errors in the search strategy). The false results were further analysed in order to establish the kind of problems being experienced. The highest percentage of errors recorded was for the incorrect use of Boolean operators as described in the help facilities of the database for a specific level of the search interface, and for the incorrect use of the broad discipline entry

box. The analysis pertaining to the complexity of the Easy Interface indicated that too many options or entry boxes to complete on a search interface may complicate decision making for end-users. As a result of this study it was decided to develop a Quick Search Interface for end-users, where the search logic is handled by choices in a language considered more appropriate for end-users.

The highest percentage of errors recorded in the 1999 study for both search interfaces was in the formulation of search strategies and the use of Boolean logic. If the information needed requires high recall and precision (as in the verification of research projects) this type of error can be disastrous. A new Quick Search Interface was therefore introduced. With the improved design, guidance is provided with on-screen brief help, general help buttons explaining Boolean searching in depth and online training material. The effectiveness of these improvements should now be monitored again, along with an assessment of the user-friendliness of the system. Screen dumps from the interfaces are included in sections 4 and 5.

3 Background on transaction log analysis

Before dealing with the actual study of the transaction log files, transaction log analysis will be briefly explained as a core quantitative method for collecting data on database and web information searching behaviour.

3.1 Definition

A number of definitions of transaction log analysis can be found in the subject literature. Nicholas et al (2002:65) explain log files as "... machine-generated records of user activity. The actual information collected by the logs depends partly on the software used and how the server was configured". Peters (1993) (as quoted by Griffiths et al 2002) defines transaction log analysis as: "The study of electronic recorded interactions between on-line information retrieval systems and the persons who search for the information found in those systems." According to Blecic et al (as quoted by Griffiths et al 2002) transaction logs should include system responses to user input to qualify as true transaction logging systems. Sandore (1993:87) explains as follows: "Transaction logs supply unequivocal information about what a user typed while searching."

Based on these definitions we take transaction logs to be machine-generated records of users' activities when using electronic systems such as electronic information retrieval system (e.g. CD-ROM databases, library catalogues, websites, intranets and web-based text databases). Activities may include information searching (e.g. the use of search terms, the use of Boolean operators), information seeking (e.g. the selection of information sources to search), the capturing of information (e.g. printing and downloading), and the beginning and ending of search sessions. The actual information collected by the logs depends partly on the software used and how the server was configured. Such information may include the length of search sessions, search queries, navigational options used (e.g. keystrokes and the use of the back button), and the system's responses.

3.2 Purpose and value of transaction logs

Transaction logs have often been appreciated for their strategic value (e.g. as pointed out by Kaske 1993; Sandore 1993). They have value for database vendors, database producers, system designers/developers, website owners, library and information professionals, information literacy trainers/educators and researchers. In

this study they will be used to enhance the CCRPD design and help facilities of the CCRPD.

Some of the benefits of transaction logs are that it is easy to collect data unobtrusively, and that data can be analysed at a later stage. Different types of statistics can be collected (e.g. the navigation moves made by users [Choo et al 1999, 2000a, 2000b], number of visits to a website such as the Nexus database system, visitors to a website, general information seeking/searching behaviour, information channels used [e.g. specific search engines, websites, portals or databases], and the users' interaction with the system [Thelwall 2001:223]). Sandore (1993:91) declares: "Transaction log analysis reveals repetitive problems in patterns of searching that should be addressed in bibliographic instruction workshops as well as in future interface designs."

Further benefits of transaction log analysis include the following:

- Can help to identify areas for database maintenance (e.g. authority files if users use different forms of a name, or spelling). Failed searches for common terms may also reveal data entry errors in the bibliographic records.
- Can yield information on the types of semantic relationships that users posit among terms and headings, which can add to the body of cataloguing and classification research. Subject access can thus be improved.
- Can be used to test the efficacy of changes to the system.
- Can determine user preferences for experimental changes.
- Can be used to anticipate the evolution of system use and demands. Shifts in users' searching behaviour over time can be picked up.
- Can support the development and improvement of the user interface.
- Can act as a decision-making tool for networks and consortia. Sandore (1993:93) writes: "In a network setting aggregate statistics reveal both similarities and differences in user searching patterns among the various institutions that comprise the network."
- Can be used to offer feedback to users on their use of the information retrieval system (e.g. the database), which is especially important when information searching is part of their professional activities.
- Marketing strategies can be developed for unutilized categories of the system and thus to draw new users. This also applies to underutilized features offered by the system.

Although the data collected through transaction logs (quantitative data) is very useful, it does not offer insight into qualitative aspects such as users' preferences, affective experiences and rationale for choices. To really draw benefit from the data collected by transaction logs, we should combine the measurement of the data (quantitative measurements) with an understanding of the rationale behind these measurements (qualitative data). This will be addressed in more detail under the heading for recommendations. Griffiths et al (2002) have the following to say about transaction logs: "Whilst transaction log analysis has considerable value as a data collection method, it has its limitations and it is best used in conjunction with a method which captures data regarding users' real information needs, comments and reactions whilst using a system and satisfaction with the system." We should therefore also note the limitations for transaction logs, for example:

- They present only snapshots of the actions of a particular set of users.
- User groups are often undefined, without distinction between novice users and information intermediaries trained to use a system. This is one aspect that we will in particular address in the second phase of this project.

- Users' levels of information literacy, education and experience with the system or the subject domain of the search strategy are not indicated.
- Reasons for the search or search strategies are not indicated.
- Users' beliefs about the information retrieval system that are logical preconditions for a search, such as that the system is a possible place to find what is wanted, are not explored.
- Users' understanding of fundamental aspects regarding multiple discourses on a topic over time within a domain of knowledge is not considered (Iivonen & Sonnewald 1998).
- Users' linguistic expressive ability is not revealed.
- The social aspects of search behaviour are not revealed (Kurth 1993).
- Transaction logs sometimes does not correlate with the users' observations of their behaviour. According to Kurth (1993) logs may show results from the point of view of the system, but may not accurately capture the users' experience and perceptions. They can help to identify only certain types of errors.

The limitations of this study should also be noted:

- Incorrect searches were not verified on the system to establish 'best practices'.
- Searchers' reasons for particular actions and choices were not determined (e.g. why do searchers use L2 and L1 in the Quick Search Interface with exactly the same words?).
- Searchers' preferences for a specific search interface were not determined.
- Only the data of the Quick Search Interface was analysed to establish whether searches were corrected on this specific interface. This is a limitation because not all interfaces were monitored.

4 Methodology of the study

Large et al (1999:43) distinguishes between information searching and information browsing. The Nexus Database System's search interfaces all support *information searching*. It is therefore assumed that for information searching the users of the CCRPD are at least able to define their information need sufficiently well to proceed with an information search strategy.

The Star software¹ developed by Cuadra was used for the development of the Nexus Database System's applications published at <http://www.nrf.ac.za/nexus>. An application for statistics was written in the Star software that stores the transactions according to three record types in the log files named as searches.all, zero.all and report.all files on the web server of all the databases of the Nexus Database System. The searches.all file contains the search strategies, hits and dates of searches. The zero.all file contains the search strategies, results of the zero hits and dates of searches. The report.all file contains the user id and transactions pertaining to the display, selection and printing of records. The search.all and the zero.all files were transferred electronically into the statistics database. It is important to note that this application was written and programmed by Cuadra to protect the privacy of users, thus no search strategy can be linked to a user.

For this study a special two-column report was written to extract the records from the statistics database that stores the search strategies, dates, and results according to

¹ Cuadra Associates, Inc. Star. Version 3.9. 11835 West Olympic Blvd., Suite 855, Los Angeles, CA 90064.

the specific search lines as displayed in the Quick and Advanced Search Interfaces of the CCRD. Two search interfaces were analysed, namely the Quick Search Interface (introduced after the previous study) and the Advanced Search Interface (adapted after the previous study). For this study a database was also developed in Star to store the search strategy data from the transaction log files with a view to longitudinal studies that will ensure the replicability of this study, and that can be used for long-term system improvement.

A number of metrics for transaction log analysis has been mentioned in the subject literature (e.g. search terms, requests, queries, unique queries, distinct queries, search sessions, accuracy, search efficiency, weighted traversal effectiveness and efficiency scores [Bilal 2000]). We decided to focus on the searches as such (including the search terms and the formulation of search strategies). A search was interpreted as an action or set of actions that led the database system to respond with a result number which may be either zero or any number of records.

Searches for the two interfaces were analysed to assess the following:

- Effectiveness of the assistance provided via the brief help facility on the search screens.
- Correct usage of the three search lines (Line 1, Line 2 and Line 3) of the Quick Search Interface to determine searchers' understanding thereof and their ability to successfully complete more complex searches.
- Searchers' understanding of the complexity of Boolean operators and the logical combination of search terms that must be used with the Advanced Search Interface.
- Correct use of search features (e.g. truncation, brackets [nesting of search terms], Boolean operators, and the use of limitations such as date, discipline, current or completed options).
- Differences in the length of search strategies between the two search interfaces (number of search terms).
- Use of synonyms and related search terms.
- Frequency of usage of the limitation options offered for the Advanced Search Interface.

The sample range was made from 1 to 28 February 2002, which constitutes a peak time for searching on the CCRPD. Four separate text files were written for the zero results and search results records of both interfaces. A record was created for each executed search strategy with its search results (zero or a number of records). These text files were imported into MS Excel for analysis and coding and named:

- QuickZero.xls with a sample size of 319 records
- QuickResults.xls with a sample size of 933 records
- AdvancedZero.xls with a sample size of 1029 records
- AdvancedResults.xls with a sample size of 912 records

In total 3193 records were analysed.

The interpretation of zero results was the same as for the 1999 study, namely true zero results (i.e. there are no hits for the search strategy) and false zero results (i.e. the result is zero, because of errors in the search strategy). The false results were further analysed in order to establish the kind of problems. Records with results but incorrect search strategies and use of the search features were also analysed. The purpose of the analysis was to determine the user-friendliness of the adapted and new search interfaces, the nature of search strategies, and frequently occurring

errors. The aim is to further improve the system to support easy and effective information retrieval, and to assess the impact of previous changes.

The data was analysed and coded according to a number of categories reflecting the features offered by the Quick Search Interface and the Advanced Search Interface. The following categories were identified:

Analysis of errors	Code or search on statistics database
<i>Step 1: All four files are coded either Zero (i.e. the result is false, because of inaccuracies in the search strategy formulation, etc.) or One (i.e. the result is true).</i> <i>Step 2: All records coded as Zero were further analysed and coded according to the categories listed in sections A and B).</i>	Column C (Coding as done in Excel) 0 or 1
<i>QuickZero.xls, AdvanceZero.xls, QuickResults.xls and AdvanceResults.xls</i>	
Section A. Category: Errors in the topic/keyword search lines	Column D (Excel)
➤ Boolean operators used incorrectly according to the specific interface requirements	1
➤ Use of phrases ²	2
➤ Spelling mistakes or hyphenation incorrectly used	3
➤ Truncation used incorrectly (mainly the Quick Search Interface where an * was used)	4
Section B. Category: Errors in the limitation section	5
➤ Incorrect entering of data in date field e.g. 20002	As a group
➤ Only one field entered and not the set of dates from To ...	
➤ Incorrect use of limitation by broad discipline (mainly without the * as truncation)	
Complexity of searches	
<i>QuickResults.xls and AdvanceResults.xls files</i>	
Section C. Category: Conceptual trends	Column E
➤ Number of records including synonyms and related terms	1
➤ Number of records using the truncation option	2
➤ Number of search terms used	Actual number of words
➤ Trend analysis of search patterns regarding narrower or broader searches	
➤ Evaluation of search lines 1, 2, and 3 of the Quick Search Interface (Abbreviated as L1, L2 and L3)	Both L1 and L2

² The two interfaces handle phrase searching differently. The Quick Search Interface allows searches in different search lines. Search line one allows for single words separated with commas to allow for searching for synonyms or related terms (e.g. libraries, museums, archives), the searching of terms combined with Boolean operators (e.g. boys OR girls). Search line two allows for concept searching (e.g. library buildings) – thus the searching of terms combined with Boolean operators (e.g. library AND buildings). Search line three allows for exact phrase searching (e.g. trade unions), thus searching terms combined with proximity operators (e.g. trade ADJ unions). The Advanced Search Interface requires that the relationship between terms be indicated with either Boolean operators or proximity operators (e.g. affirmative ADJ action) whereas the Quick Search Interface allows searching without the use of operators.

	completed, L1 commas not used
➤ Trend analysis of the use of the options next to the keyword entry lines in the Advanced Search Interface. These include searching in the title, abstracts or all the basic indexes (i.e. subject related fields)	Index on statistics database
Section D. Limitation frequency	
➤ Number of records with keyword transactions limited by years	Search on statistics database
➤ Number of records with keyword transactions limited by broad subjects (disciplines)	Search on statistics database

Table 1: Categories for analysis

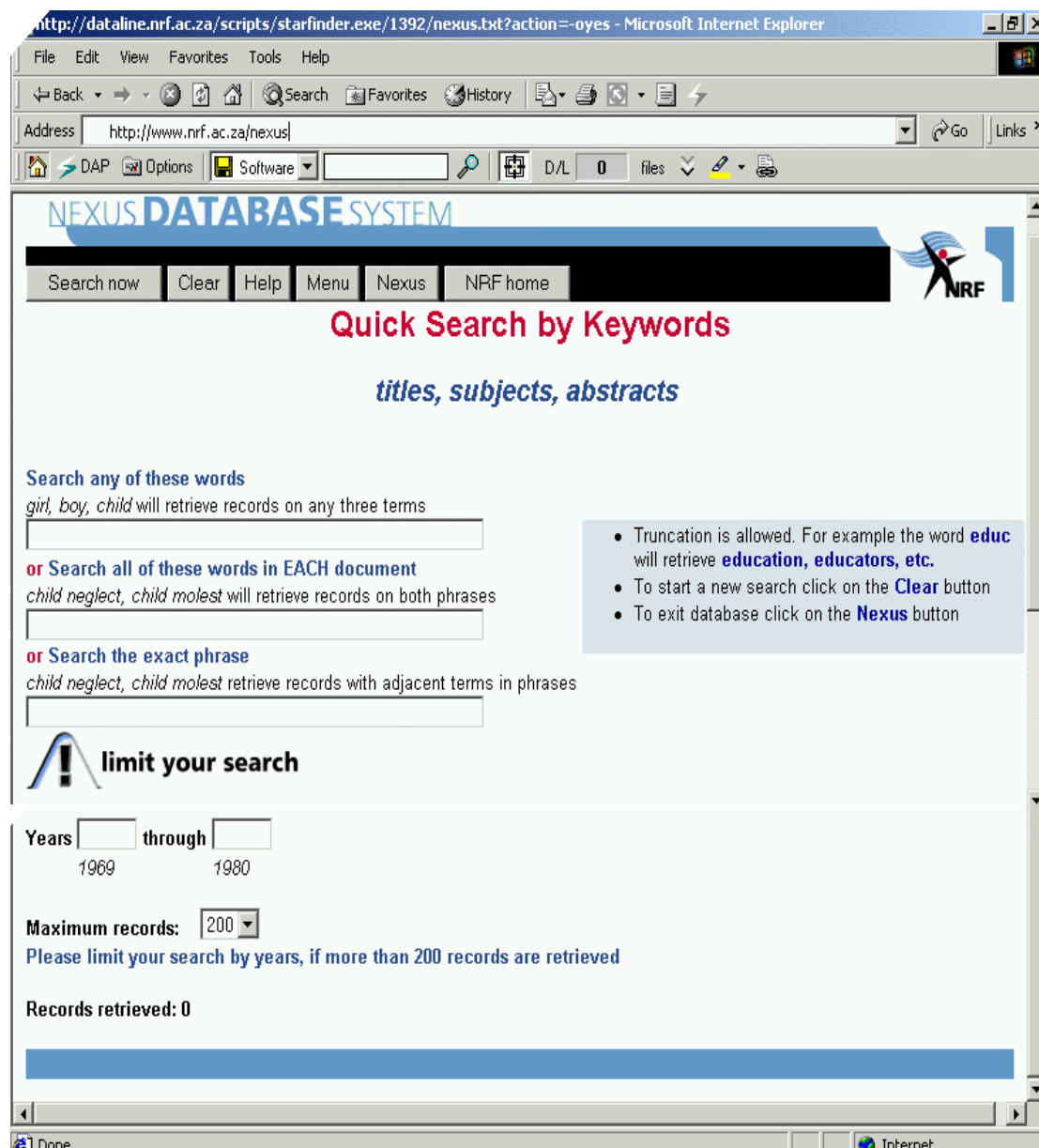
Sections A and B concern the analysis of errors in the zero as well as the results files for the two interfaces. Categories were identified for common errors in the keyword and topic searching, as well as for errors on limitation of search results. Sections C and D concern results from both interfaces that have records, although the search strategies are either wrong or incomplete. These sections concern the complexity of the search strategies, and include conceptualisation, and the use of limitation options.

5 Data analysis of the Quick Search Interface

The Quick Search Interface was designed for end-users and intermediaries respectively as illustrated below.³

³ An end-user is a person who has not received formal training in search formulation and strategies. An intermediary is an information consultant/librarian who has received extensive training in conceptualizing search strategies and formulating a search statement.

Figure 1: Quick Search Interface, 2002



5.1 Analysis of errors: Sections A and B

A total of 1147 searches were analysed, with a total number of 529 errors. Searches where communication breakdowns and institutional access problems were suspected, were excluded. These searches were coded as 8 and 9 respectively in the table, and they were not included in the number of searches analysed.

The percentage of errors and different categories of errors are depicted in Figures 2a and 2b.

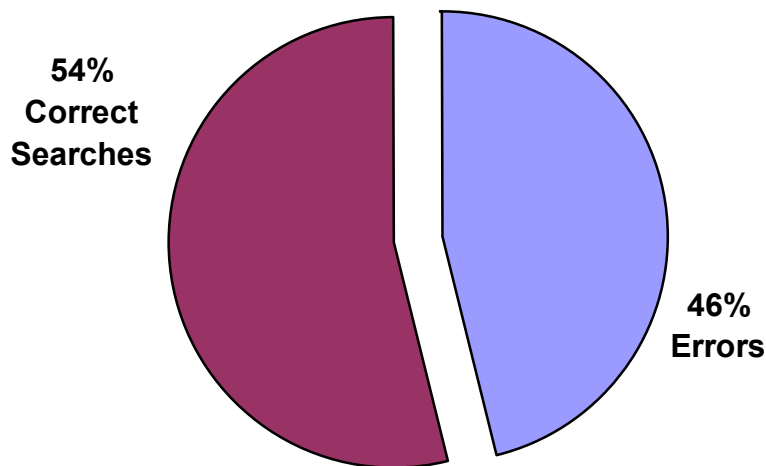


Fig. 2a: Percentage of errors

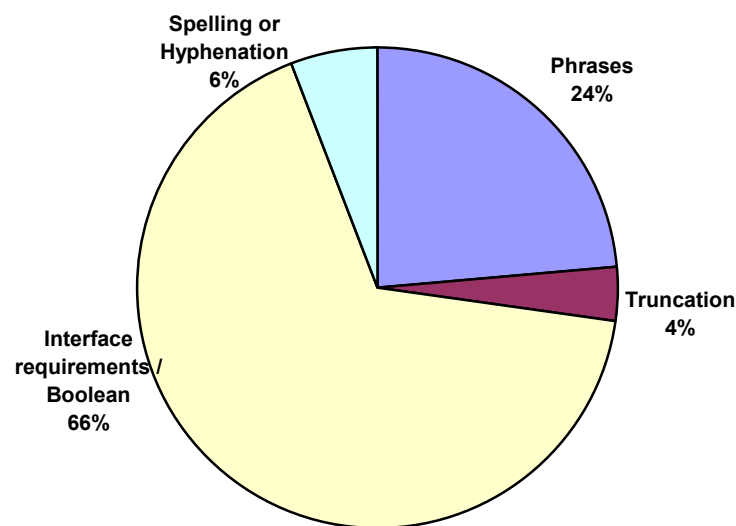


Fig. 2b: Categories of errors

The percentage of errors is quite concerning in the view of the purpose of the database. The highest percentage of errors recorded for the Quick Search Interface, as shown in Figure 2b, was for incorrect use of the three search lines according to the specific requirements for this interface as well as the use of phrases rather than keywords or single concepts. Example 1a shows that a phrase was entered in search line three. Example 1b indicates the incorrect use of Boolean operators that are not required for the Quick Search Interface. Thus, the search for Example 1b should be entered in search **line one** as *transformation, restructuring*. In Example 2 all three search lines were completed with more than one error, namely the use of a slash in search line one, which is incorrect, and a rather long phrase (almost a sentence) in search line 3. The result of 214 is thus incorrect. If we assume that the topic is *career aspirations*, it would have resulted in the retrieval of 24 records when only entered into search line two as *career aspirations*.

Example 1a

--REC-- 375510
L30 2/8/2002
L3 impact communication between employer and employees
L4 1969
L5 2002
L6 {ITEMS -1:-200}
L20 0 (**Number of records retrieved**)

Example 1b

--REC-- 379730
L30 02/21/2002
L2 (transformation OR restructuring)
L3 local government
L6 {ITEMS -1:-200}
L20 1 (**Number of records retrieved**)

Example 2

--REC--378416
L30 02/09/02
L1 career/occupations
L2 career aspirations
career aspirations and perceptions of South African
L3 children
L4 1998
L5 2001
L6 {ITEMS -1:-50}
L20 214 (**Number of records retrieved**)

**5.2. Analysis of the complexity of searches in the Quick Search Interface:
Sections C and D**

A total number of 928 searches was analysed. Searches using truncation, synonyms and Boolean operators are considered to be more complex than searches on single terms or concepts. The complexity of searches often gives an indication of the effective use of search features. Only 5% of searches indicate the use of synonyms and 2% of searches make use of truncation. Example 3a is a good example of a search that was corrected by the searcher. Even so, a better result would be obtained by using the correct truncation as well. If truncation was used for emerging markets, e.g. *L2 emerg market*, 152 records would have been retrieved. Example 3b would have retrieved 14 records if truncation was used.

Example 3a

--REC-- 379056
L30 02/14/2002
L2 characteristics of emerging markets
L4 1990
L5 2000
L6 {ITEMS -1:-10}
L20 1 (**Number of records retrieved**)
--REC-- 379057
L30 02/14/2002

L2 emerging markets
 L4 1990
 L5 2000
 L6 {ITEMS -1:-10}
 L20 22 (**Number of records retrieved**)

Example 3b

--REC-- 378916
 L30 02/13/2002
 L2 Crime gauteng
 L6 {ITEMS -1:-200}
 L20 12 (**Number of records retrieved**)

Searches on the Quick Search Interface showed an average of 2.57 words per search. This can, however, mostly be ascribed to the incorrect use of long phrases that constitute 24% of the errors. Limitation by years resulted in 37%. Limitation by years is therefore used to a limited extent, which might be explained by gathering more information on the purpose of the searches and the specific needs for the information.

It appears that the search topics were not analysed appropriately to construct a search strategy logically, and that it was therefore also not possible for the users to select the correct interface to be used⁴. The correct interface for example 4 below .would have been the Advanced Search Interface with the search strategy as follows: (information ADJ technolog*) AND (learn* OR teach*).

Example 4

--REC-- 378961
 L30 02/14/2002
 L3 Information technology in learning and teaching
 L4 1999
 L5 2002
 L6 {ITEMS -1:-200}
 L20 1 (**Number of records retrieved**)

Although 21,4 % of the searches were corrected, the results show that in many of these, the compilation of the search strategy needs further refinement and that the use of synonyms or related terms should be included to optimize the results. This would ensure that all relevant references were retrieved.

Example 5

--REC--
 - 368582
 L30 4/2/2002
 L2 Smoking
 L6 {ITEMS -1:-200}
 L20 73 (**Number of records retrieved**)

By using related terms (in Example 5), for example entering *smoke*, *smoking*, *tobacco*, *cigar*, *cigarette* in L1, 241 records would be retrieved.

⁴ It therefore seems that conceptual analysis of the information need might be a serious problem that should be further researched by using interviews with users for example.

From the above-mentioned data and examples, it appears that users should be trained in conceptualization and the selection of suitable search strategies. Before embarking on this route, the results should, however, be verified by follow-up methods.

6. Data analysis of the Advanced Search Interface

The data for the Advanced Search Interface was analysed in a similar fashion. The interface is depicted in Figure 3.

Figure 3: Advanced Search Interface, 2002

NEXUS DATABASE SYSTEM

Search now Clear Help Professional search Menu Nexus NRF home

Advanced Search by Keywords

1. To search "**Keywords**"

- Select significant words
- Separate these words with Boolean or Proximity operators e.g. **child* adj abuse*** The use of a wildcard "*" is allowed.
- Use brackets to combine your search terms logically, e.g. **(child* ADJ abuse*) AND ((Pretoria OR Durban Or (Cape ADJ Town))**
- You may use both lines to enter your search. Use the **Radio buttons** to combine the searches logically.

GO TO

Quick SEARCH

Advanced SEARCH

When you have entered your search click on

Keywords
i.e. titles, subjects

OR AND ADJ NEAR W/O

limit your search

All current projects on your topic

All completed projects on your topic

Current: Year from through

Completed: Year from through

Discipline(s)

e.g. **024* or 022*** (click on HELP for a list of codes).

When you have entered your search click on

Maximum records:

Please limit your search, if more than 200 records are retrieved

Records retrieved: 0

6.1. Analyses of errors: results on Sections A and B

A total of 1780 searches were analysed, with a result of 641 errors.

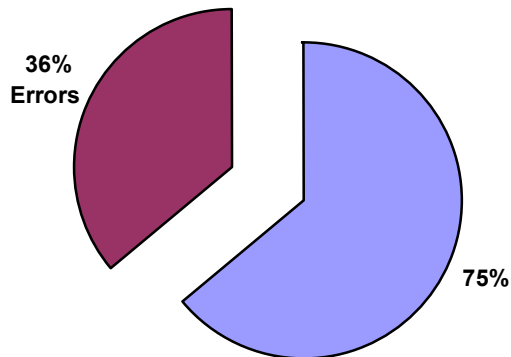


Fig. 4a: Percentage of errors

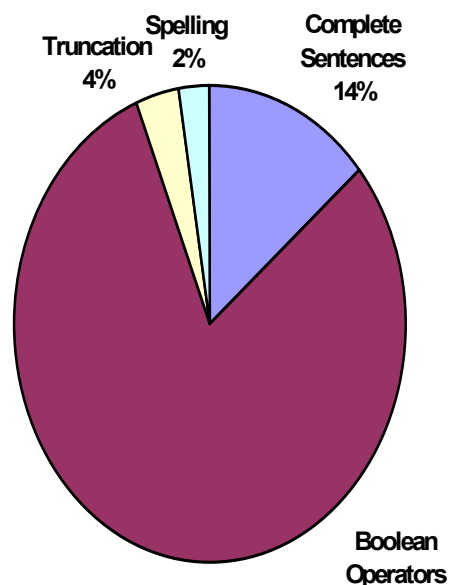


Fig. 4b: Categories of errors

The percentage of errors is less than that for the Quick Search Interface. The highest percentage of errors recorded for the Advanced Search Interface, as reflected in Figure 4b, was for the incorrect use or non-use of Boolean operators between terms as described in the help facilities. The use of phrases also caused problems. Example 6, for instance requires a Boolean or proximity operator between the two terms in L1 *trade unions*, for example *trade AND unions* or *trade ADJ unions*.

Example 6

L30	2/1/2002
L1	trade unions
L2	BI
L3	OR
L5	BI

L12 {ITEMS -1:-200}
 L20 0 (Number of records retrieved)

Example 7 below requires the Boolean operator OR, for example L1 *gay** OR *homosexual** OR *lesbian**, and truncation with the asterisk "*" would have ensured that all relevant records are retrieved. Once again the reason why there are fewer errors can be contributed to a number of things, such as a clearer interface or more sophisticated searchers. Follow-up interviews with searchers should, however, provide us with more insight.

Example 7

L30 2/4/2002
 L1 Gay
 L2 BI
 L3 OR
 "L4 homosexual, lesbian"
 L5 BI
 CC=CURRENT* or ONGOING* or INDEFINITE* OR
 L6 CC=COMPLETED*
 L11 022*
 L12 {ITEMS -1:-200}
 L20 2 (number of records retrieved)

6.2 Analysis of the complexity of searches: Sections C and D

The use of synonyms was very low. They were used only for 45 out of 908 (5%) searches, and the use of truncation was 30%. This is a major concern when some of the benefits of the CCRPD are to avoid duplication in research activities and the identification of new developments, gaps and trends in the study fields. This also has an impact on the successful retrieval of ALL relevant records for a particular topic.

The use of limitation by years and disciplines is low and resulted in 10% for years, and 14% for disciplines. It might be that searchers do not know how to use the limitation options, or there is little need for their use. Furthermore, the discipline field indicates 24% errors (see Fig. 5). The most errors were for not using truncation, namely the asterisk "*" as indicated in the "help" line on the search screen below the entry box (see Fig. 3). The non-use of the asterisk for truncation is illustrated in examples 8a and 8b.

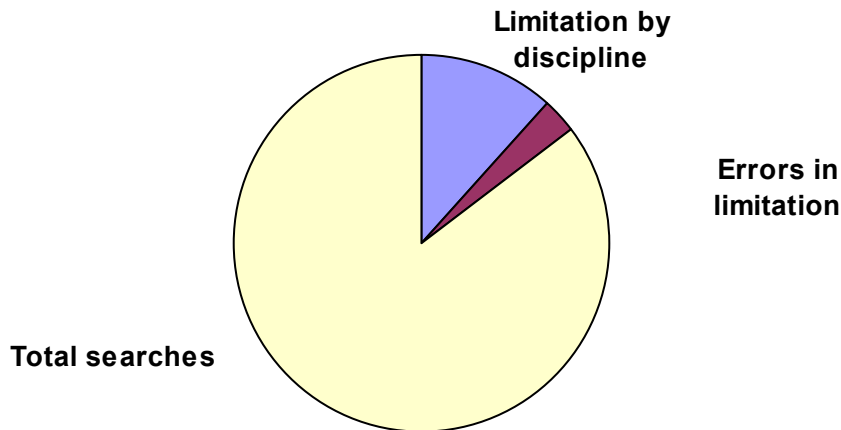


Fig. 5: Limitation by disciplines and errors

Example 8a

L30 2/8/2002
L1 Music
L2 BI
L3 AND
L4 outcomes-based education
L5 BI
L6 CC=CURRENT* or ONGOING* or INDEFINITE*
L11 **017**
L12 {ITEMS -1:-200}
L20 0 (Number of records retrieved)

Example 8b

L30 2/28/2002
L1 Role conflict
L2 BI
L3 OR
L5 BI
L6 CC=CURRENT* or ONGOING* or INDEFINITE* OR CC=COMPLETED*
L7 1990
L8 2002
L9 1990
L10 2002
L11 **24**
L12 {ITEMS -1:-200}
L20 0 (Number of records retrieved)

The number of words used in the Advanced Search Interface was 1.9 words per search line. The provision of two search lines for this interface requires further investigation in order to establish the need for its complexity. The use of the option to execute a search in the title field, abstract field, all the basic index fields (i.e. all the subject-related fields), et cetera, is indicated in Figure 6.

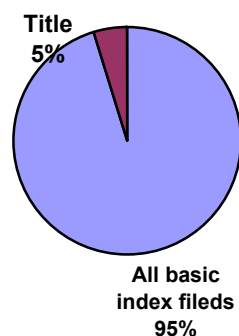


Fig. 6: Use of option box next to keyword entry line

7 General observations of errors that have occurred frequently

It was noted with surprise that sometimes a search is repeated three to four times with exactly the same spelling mistake or incorrect phrase. Very specific searches (narrow searches) were also seldom executed as a broader search. It also appears that if results should be interpreted with care: system zero results for a particular day, namely 07/02/2002 (with no theoretical problems with the search terms or search strategies), indicated, for example that there was probably a problem at a specific institution because the QuickResult.xls file did, in fact, contain successful searches.

8 Interpretation and recommendations

Before interpreting the data, we took note of Kurth's (1993:102) warning: "In as much as researchers' relationship with institutions may affect the research results they choose to report, so may researchers' relationships with their own research agendas affect the conclusions they draw from those results. Results of analyses that identify problems in the transactions between systems and users do not automatically imply the remedies to those problems. When a researcher concludes from a transaction log analysis that system changes or user education programs will address the searching problems uncovered by the analysis such conclusions may reflect the researcher's desire to change the system or educate the users more than the researcher's considered interpretation of the study's results. Verifying the conclusions drawn from the results of transaction log analyses and the implications of those results requires a methodology as precise as that of transaction log analysis itself."

This paper is merely a preliminary report to highlight the ease with which valuable data can be collected through transaction logs. Among other things it indicates that there are problems with the use of the two interfaces, and especially the Quick Search Interface. At this stage, we can only speculate on the cause of the problems, be it the design of the interface or searchers with little understanding of the conceptualisation of information needs or inadequate or no training in the use of search features such as Boolean operators.

At this stage all users are treated the same. We should, however, distinguish between librarians and information specialists, searches by academics and researchers, and searches by students, when improving the survey and implementing supplementary methods of data collection, such as interviews and even focus group interviews.

9 Conclusion

Transaction logs offer very valuable data for establishing problem areas, as we have seen in the case with the use of the two Nexus interfaces for the CCRPD. The data should, however, be interpreted with care. Instead of rushing into changes, we should use the data should be used to develop instruments (such as interview questionnaires) for verifying the data, and to ask the right questions during such interviews in order to collect data that can be used for meaningful changes to the user interfaces and/or training programmes.

Two suggestions for implementation at this stage are:

- The use of one search line for the Quick Search Interface. This should be monitored over a brief period. It will be easy to implement.

- Examples of hyphenated search requirements for the different interfaces should be provided on the search screens.

References

- Bilal, D. 2000. Children's use of the Yahoo!igans! Web search engine. Part 1. Cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of the American Society for Information Science*, 51(7):646-665.
- Choo, C.W., Detlor, B. & Turnbull, D. 1999. Information seeking on the web: an integrated model of browsing and searching. *ASIS Annual Meeting*. [Online]. <http://choo.fis.utoronto.ca/fis/respub/asis99/>. Accessed 14 March 2002.
- Choo, C.W., Detlor, B. & Turnbull, D. 2000a. *Working on the Web: an empirical model of web use*. Paper presented at the HICSS 33 (Hawaii International Conference on Systems Science), 4–7 January 2000. [Online]. <http://choo.fis.utoronto.ca/fis/respub/HICSS/default.html/>. Accessed 14 March 2002.
- Choo, C.W., Detlor, B. & Turnbull, D. 2000b. *Web work: information seeking and knowledge work on the world wide web*. Dordrecht: Kluwer Academic.
- Cooper, M.D. 2001. Usage patterns of a Web-based library catalog. *Journal of the American Society for Information Science and Technology*, 52(2):137-148.
- Fourie, I. 2002. A review of web information-seeking/searching studies (2000 – 2002): implications for research in the South African context. In *Progress in Library and Information Science in Southern Africa: proceedings of the second biennial DISSAnet Conference (PROLISSA conference, 24-25 October, Pretoria)*. Edited by T. Bothma & A. Kaniki. Pretoria: Infuse: 49-75. [Also available online: <http://www.dissanet.com>]
- Griffiths, J.R., Hartley, R.J. & Willson, J.P. 2002. An improved method of studying user-system interaction by combining transaction log analysis and protocol analysis. *Information Research*, 7(4) [Online]. <http://InformationR.net/ir/7-4/paper139.html>. Accessed 3 September 2002.
- Iivonen, M. & Sonnenwald, D.H. 1998. From translation to navigation of different discourses: a model of search term selection during the pre-online stage of the search process. *Journal of the American Society for Information Science*, 49(4):312-326.
- Jansen, B.J. & Pooch, U. 2001. A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52(3):235-246.
- Kaske, N.K. 1993. Research methodologies and transaction log analysis: issues, questions, and a proposed model. *Library Hi Tech*, 11(2):70-86.
- Kurth, M. 1993. The limits and limitations of transaction log analysis. *Library Hi Tech*, 11(2):98-104.
- Large, A., Tedd, L.A. & Hartley, R.J. 1999. *Information seeking in the online age : principles and practice*. London, Bowker-Sauer.

Marchionini, G. 1995. *Information seeking in electronic environments*. Cambridge, UK: Cambridge University Press.

Nicholas, D., Huntington, P., Lievesley, N. & Withey, R. 1999. Cracking the code: web log analysis. *Online & CD-ROM review*, 23(5):263-269.

Nicholas, D., Huntington, P. & Williams, P. 2001. Establishing metrics for the evaluation of touch screen kiosks. *Journal of Information Science*, 27(2):61-72.

Nicholas, D., Huntington, P. & Williams, P. 2002. Evaluating metrics for comparing the use of web sites: a case study of two consumer health web sites. *Journal of Information Science*, 28(1): 63-75.

Peters, T. 1993. The history and development of transaction log analysis. *Library Hi Tech*, 42(1):41-66.

Sandore, B. 1993. Applying the results of transaction log analysis. *Library Hi Tech*, 11(2):87-97.

Spink, A., Wolfram, D., Jansen, M.B.J. & Saracevic, T. 2001. Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226-234.

Thelwall, M. 2001. Web log file analysis: backlinks and queries. *Aslib proceedings*, 53(6):217-223.

Van der Berg, H. 2000. An analysis of search strategy trends in order to enhance the web search interface of the Nexus database system. CRIS2000 Conference, 25-27 May 2000, Espoo, Finland. [Available at <http://www.cordis.lu/cris2000/src/product.htm>.] Accessed 21/02/2003.