Approaches to Understanding the Searching Behavior of CrossFire Users

Frank Cooke and Nikolai Kopelev

GlaxoSmithKline, 709 Swedeland Road, P.O. Box 1539, King of Prussia, Pennsylvania 19406-0939

Helen Schofield*

Chemistry Department, UMIST, P.O. Box 88, Sackville Street, Manchester M60 1QD, UK

Geraldine Boyce

GlaxoSmithKline, New Frontiers Science Park, Harlow, Essex, CM19 5AW, UK

Sean Dunne

MIMAS, Manchester Computing, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Received February 28, 2002

Due to the high costs of purchasing, supporting and training users of desktop chemical information systems, it is important to understand users' behavior in order that deficiencies in their search efficiency and effectiveness can be identified and addressed. CrossFire generates comprehensive log files that can be examined to determine the nature of search activity. At GSK (GlaxoSmithKline) a Log File Parser, CrossParse, has been developed in Visual Basic that enables analysis by individual user name, groups of users or the whole user population. Log files can be analyzed for occasions when specific structural features are built, specific types of search are done and how the results are manipulated. CrossParse produces output that can be saved and analyzed within Microsoft Excel. It also allows determination of numbers of active concurrent users on the CrossFire system. CrossParse has been used at GSK (ex-SmithKline Beecham sites) to examine the search behavior of medicinal and synthetic chemists. Additionally, it has been used by MIMAS (Manchester Information and Associated Services) to compare the search behavior of trained and untrained users in the higher education community and to identify any areas where improvements to training can be made. Use of CrossParse in both organizations has allowed identification of areas where users may have difficulties using CrossFire. This will provide valuable feedback to MDL Inc., the authors of the CrossFire application, and guide them in enhancing CrossFire.

INTRODUCTION

Desktop chemical information systems targeted toward end-users are generally quite expensive. This expense is derived from several sources including the original purchase of the application, the rollout to the end-users' desktops, training, day to day support, and the often forgotten soft costs associated with the purchase of internal servers, technical support staff for the servers and scheduled updates of the application database. It is important that the benefit/cost ratio be maximized for such investments. One component, necessary to achieve this, is a good understanding of how effectively the user population utilizes the application.

There have been a number of studies performed with the expectation of gaining a better understanding of end-users' searching behavior. Many of these have been associated with Online Public Access Catalog systems (OPACs). Often, these studies have been performed via two distinct mechanisms: (1) subjective—surveys that are completed by the end-users that describe their perception of their searching behavior and (2) objective—transaction log analysis.

Most modern day information access applications produce transaction logs. These are collections of 'activities' performed by the end-user and often include data such as the following: date, time, search string, navigation operation, answer display and an end-user identifier. Frequently, these transaction logs are analyzed manually by first printing them and then by visual inspection. This is the more subjective but insightful approach compared to relying on programmatic analysis only. However, this can be very time-consuming and does not lend itself well to the analysis of large transaction logs or repeated analysis on a periodic basis. By combining the analysis of the transaction logs with end-user surveys often a more complete picture of searching behavior can be achieved.

Some transaction-log studies have been performed on OPAC systems to gather information on which type of information is being accessed so that more targeted collection development might be achieved.[1] Other studies have investigated the effect of additional searching aids, e.g. flip charts next to the terminal or additional online help.[2] This help might be in the form of a separate Web page or built-in application help in the typical Windows/Help menu format. Similar studies have looked at user behavior as it relates to application design in order to optimize ease of use, remove

* Corresponding author phone: +44-161-200-4468; e-mail: helen.schofield@umist.ac.uk.

10.1021/ci0202638 CCC: $22.00 © 2002 American Chemical Society
Published on Web 08/16/2002

points of confusion and help guide the user to desired result.[3−6]

Numerous papers have been written about the search behavior of sets of users and investigations around the types of searches being performed. Most of these are aimed at optimizing or customizing the user interface and the methodology behind a particular search system to enhance the user interaction and allow better retrieval and precision.[7−14] Recent work has also focused on users' behavior when searching the Web.[15]

Often, in the area of chemistry, the most important search criteria are buried within the chemist's language of communication, i.e., the chemical structure. The CrossFire application offers very rich transaction logs that incorporate complete specifications of the search in the language of the chemist. These logs can be extremely large and do not lend themselves to visual inspection. For this reason a parser application, we are calling CrossParse, was written in Visual Basic to attempt to understand how effectively end-users are using CrossFire.

Programmatic analysis of transaction logs is not new. One example is the analysis of a library transaction log via a Fortran program to analyze for search types and correlating this with particular times within a semester.[16] Some useful background on the challenges and opportunities for transaction logging and its analysis can be found in the work by Jones, Gatford and Walker.[17] However, to our knowledge this type of analysis has never been undertaken for a chemistry application using chemistry search features.

MIMAS (Manchester Information and Associated Services) provides many training classes throughout the year to the UK, Scandinavian and Irish higher-education communities, and CrossParse offers a means to view a 'before and after' picture of how users are employing the application and specifically those features or functionalities that have been addressed during a prior training event. Periodic use of CrossParse allows the determination of whether the 'training effect' has a long lasting effect versus an immediate but unsustainable effect.

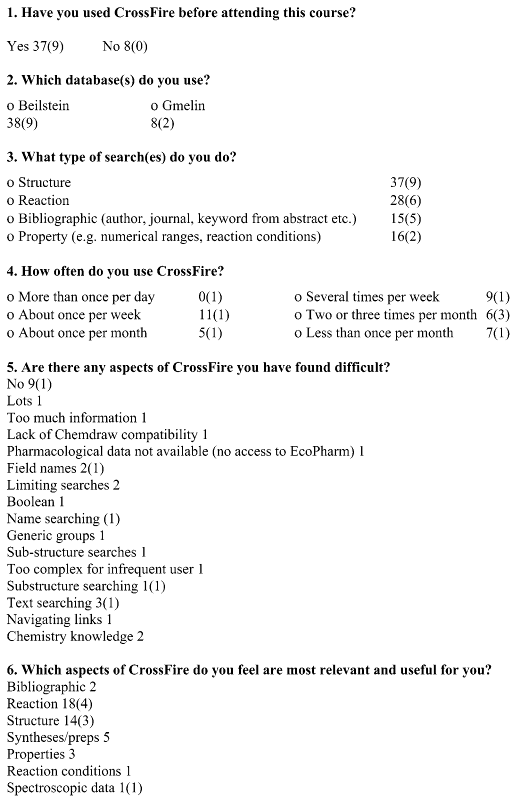
Comparisons have also been made by other authors between trained and untrained user search behavior,[18−21] but these have usually been undertaken by use of questionnaires or direct observations of a relatively small number of searchers and therefore tend to be rather subjective. CrossParse allows a more quantitative approach to this type of analysis.

One of the political and social issues not often addressed in works of this type is the 'big brother' syndrome, i.e., a higher authority is watching over you 'for your own good'. It should be made clear to users before embarking on such an endeavor that the desired outcome of such a work is to optimize training effectiveness and their productivity and not to point at 'troublesome' individuals for specific remedial action! It should be noted that at no stage during this research were individuals identified.

MIMAS CROSSFIRE SERVICE

Background. CrossFire (Beilstein and Gmelin)[22] are provided to the UK higher education community via MIMAS[23,24] at the University of Manchester, a JISC[25] (Joint Information Systems Committee) supported national data center, which provides the UK higher education, further education and research communities with networked access to key data and information resources to support teaching, learning and research across a wide range of disciplines. Agreements have been made with consortia of universities in Sweden, Denmark and Finland, bringing the total number of sites which access the CrossFire service via MIMAS to 85. At present there are about 4500 active users of the system. The user community is diverse, and, in addition to chemists, it also includes a number of chemical engineers, biochemists, materials scientists, physicists, even electrical engineers, business studies students and librarians (who usually do not have a background in chemistry). Additionally, users are at different stages of their studies, ranging from first-year undergraduates through postgraduates and other researchers to academic staff.

Table 1. MIMAS Questionnaire Distributed at Face-to-face Training Courses[a]



1. Have you used CrossFire before attending this course?

Yes 37(9) No 8(0)

2. Which database(s) do you use?

o Beilstein	o Gmelin
38(9)	8(2)

3. What type of search(es) do you do?

o Structure	37(9)
o Reaction	28(6)
o Bibliographic (author, journal, keyword from abstract etc.)	15(5)
o Property (e.g. numerical ranges, reaction conditions)	16(2)

4. How often do you use CrossFire?

o More than once per day	0(1)	o Several times per week	9(1)
o About once per week	11(1)	o Two or three times per month	6(3)
o About once per month	5(1)	o Less than once per month	7(1)

5. Are there any aspects of CrossFire you have found difficult?
No 9(1)
Lots 1
Too much information 1
Lack of Chemdraw compatibility 1
Pharmacological data not available (no access to EcoPharm) 1
Field names 2(1)
Limiting searches 2
Boolean 1
Name searching (1)
Generic groups 1
Sub-structure searches 1
Too complex for infrequent user 1
Substructure searching 1(1)
Text searching 3(1)
Navigating links 1
Chemistry knowledge 2

6. Which aspects of CrossFire do you feel are most relevant and useful for you?
Bibliographic 2
Reaction 18(4)
Structure 14(3)
Syntheses/preps 5
Properties 3
Reaction conditions 1
Spectroscopic data 1(1)

[a] Total number of completed questionnaires is 54, including 11 who have not received the follow-up survey. Numbers in parentheses are for the 9 respondents to the follow-up questionnaire.

CrossFire Training. The specialized nature of CrossFire ensures that some form of training is required in order to use the system to a professional level. Unlike purely text-based systems, there are a number of different aspects of CrossFire whose interfaces are not immediately intuitive to the new user. There is a continual demand for MIMAS CrossFire training courses as the MIMAS user community is transient: each year a new cohort of undergraduate and postgraduate students arrives at universities. The main attendees of MIMAS CrossFire training courses are postgraduate students. Around 1500 CrossFire users have been trained by MIMAS during the five years of the service.

Assessment of Training. Traditionally, MIMAS training courses have been evaluated by use of questionnaires completed by attendees after course attendance. This assessment evaluates the course itself, its content, delivery and the course materials, and is necessarily qualitative in nature. The questionnaire does not measure any increase in efficiency in searching resulting from the training or identify whether course attendees subsequently make use of any search features covered in the course which were not used prior to attending the course.

More thorough investigation by questionnaire is possible, and this has recently been introduced by MIMAS for its CrossFire training. This takes the form of an in-depth questionnaire distributed at the course. Attendees are invited to provide additional background information and answer a selection of questions, which are shown in Table 1 along with a summary of responses provided at training courses delivered at three universities in 2001.

Approximately one month after the course a further questionnaire is distributed by e-mail to attendees who completed the initial questionnaire and who have given their agreement to taking part in the follow-up study. The follow-up questions asked at this stage and answers provided by trained users at the three sample universities are given in Table 2.

The log-file parser developed at ex-SB (SmithKline Beecham) provides the opportunity for a more objective, semiquantitative assessment of the effectiveness of the MIMAS CrossFire training courses. This can be achieved by running the program for specific groups of users or even individual users before and after they have attended training courses. Additionally it facilitates comparison of actual user behavior with users' perceptions of the benefit provided by training.

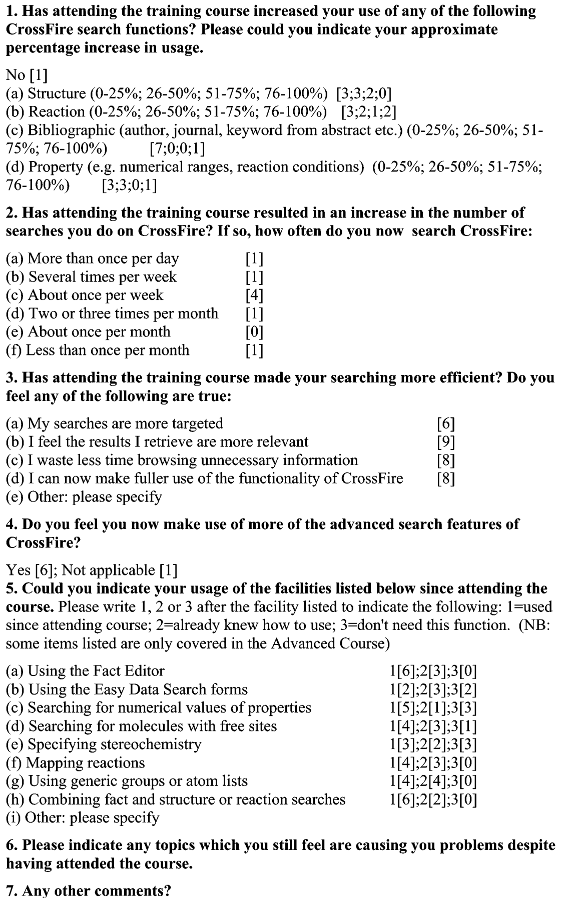
CROSSFIRE AT EX-SB

CrossFire was released to over 400 end-users at SB in October 1997. On-site hands-on training was arranged for those users able to attend the scheduled training classes. For those unable to attend separate one-on-one training or group training without hands-on experience was offered. Following the launch of CrossFire, any additional users requesting access to the system were only offered one-on-one training as the number requesting training on a periodic basis was insufficient to warrant organized classroom training on a regular basis.

A general questionnaire was prepared and used following the mass training for the initial launch. This provided feedback on the Beilstein professional training staff and the material covered in much the same way as the traditional evaluation done by MIMAS. There has been no systematic effort employed to capture the effectiveness of the subsequent one-on-one training.

SB's primary audiences for the CrossFire application are Medicinal Chemistry and Chemical Development. Chemical Development can be broadly split into two sections—Synthetic Chemistry and Analytical Chemistry. Usage by Medicinal Chemistry and Synthetic Chemistry represents greater than 95% of the usage of the system. The other ∼5% is primarily made up by the Information Management department with little use being made by the Analytical Chemistry department. Usage by the Information Management department is on behalf of those chemists that either do not feel comfortable using the application or need a complex or comprehensive search done.

Table 2. Follow-up Questionnaire Distributed by E-mail to Course Attendees Who Completed the Initial Training Questionnaire (See Table 1)[a]



1. Has attending the training course increased your use of any of the following CrossFire search functions? Please could you indicate your approximate percentage increase in usage.

No [1]
(a) Structure (0-25%; 26-50%; 51-75%; 76-100%) [3;3;2;0]
(b) Reaction (0-25%; 26-50%; 51-75%; 76-100%) [3;2;1;2]
(c) Bibliographic (author, journal, keyword from abstract etc.) (0-25%; 26-50%; 51-75%; 76-100%) [7;0;0;1]
(d) Property (e.g. numerical ranges, reaction conditions) (0-25%; 26-50%; 51-75%; 76-100%) [3;3;0;1]

2. Has attending the training course resulted in an increase in the number of searches you do on CrossFire? If so, how often do you now search CrossFire:

(a) More than once per day	[1]
(b) Several times per week	[1]
(c) About once per week	[4]
(d) Two or three times per month	[1]
(e) About once per month	[0]
(f) Less than once per month	[1]

3. Has attending the training course made your searching more efficient? Do you feel any of the following are true:

(a) My searches are more targeted	[6]
(b) I feel the results I retrieve are more relevant	[9]
(c) I waste less time browsing unnecessary information	[8]
(d) I can now make fuller use of the functionality of CrossFire	[8]
(e) Other: please specify	

4. Do you feel you now make use of more of the advanced search features of CrossFire?

Yes [6]; Not applicable [1]
5. Could you indicate your usage of the facilities listed below since attending the course. Please write 1, 2 or 3 after the facility listed to indicate the following: 1=used since attending course; 2=already knew how to use; 3=don't need this function. (NB: some items listed are only covered in the Advanced Course)

(a) Using the Fact Editor	1[6];2[3];3[0]
(b) Using the Easy Data Search forms	1[2];2[3];3[2]
(c) Searching for numerical values of properties	1[5];2[1];3[3]
(d) Searching for molecules with free sites	1[4];2[3];3[1]
(e) Specifying stereochemistry	1[3];2[2];3[3]
(f) Mapping reactions	1[4];2[3];3[0]
(g) Using generic groups or atom lists	1[4];2[4];3[0]
(h) Combining fact and structure or reaction searches	1[6];2[2];3[0]
(i) Other: please specify	

6. Please indicate any topics which you still feel are causing you problems despite having attended the course.

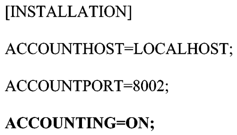
7. Any other comments?

[a] Nine responses were received (out of 45 sent). Numbers of responses are given in square brackets.

Within the new GSK organization a number of new and more effective training approaches are being developed. It is anticipated that the work described here will provide useful metrics for the effectiveness of the new training programs, currently under development, for our scientific desktop tools.

CROSSPARSE DESIGN AND FUNCTIONALITY

Most of the actions taken by CrossFire end-users can be captured to a transaction log. If the transaction log functionality has been licensed, then by default, it is named 'xfacct.log' and is stored in the XFIRE\SYS directory on the CrossFire server. When the option for logging is turned on, there are several options that determine the level of logging to be performed. In the XFIRE.INI file, located in the XFIRE\BIN directory, there should be an entry for ACCOUNTING that is set to 'ON':

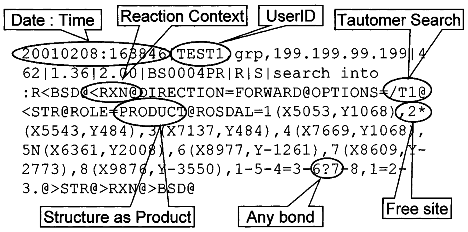


```
[INSTALLATION]

ACCOUNTHOST=LOCALHOST;

ACCOUNTPORT=8002;

ACCOUNTING=ON;
```



```
20010208:163846|TEST1,grp,199.199.99.199|4
62|1.36|2.00|BS0004PR|R|S|search into
:R<BSD@<RXN@DIRECTION=FORWARD@OPTIONS=/T1@
<STR@ROLE=PRODUCT@ROSDAL=1(X5053,Y1068),2*
(X5543,Y484),3(X7137,Y484),4(X7669,Y1068),
5N(X6361,Y2008),6(X8977,Y-1261),7(X8609,Y-
2773),8(X9876,Y-3550),1-5-4=3-6?7-8,1=2-
3.@>STR@>RXN@>BSD@
```

Figure 1. A snapshot of the ROSDAL code for a CrossFire reaction search, with some of the structure and reaction details indicated.

Table 3. CrossFire Log File Format, Indicating the Structure of the Log Files and Type of Data Recorded

column	contents
1	time/date using format: YYYYMMDD:HHMMSS
2	user name + group name + TCP/IP address (separated by commas)
3	process number
4	CPU time of the command
5	elapsed time of the command
6	CrossFire database being used
7	database context of the command (S = substance, R = reactions, C = citation, N = no context found)
8	reference character for the command type (S = structure search for substance or reaction, F = fact search, <space.> = not a searching command)
9	command or search statement

The details of the accounting are found in the XFIRE.CFG file which is located in the SYS directory. An example is shown here:

```
[Accounting]

Port=8002;

AccountCommands=search_,,send_,,dbselect,init,exit;

AccountLog=/xf4/sys/xfacct.log;
```

The accounting function can collect specific commands or all commands that the CrossFire user is sending to the CrossFire server. If the command list is empty, then all commands are logged. Using the "_" after a specific command signifies that the command should be logged but not the complete command string associated with it. Logging all commands with all details is the appropriate option for CrossParse.

Table 4. Examples of Representations of Structural Features Which Can Be Automatically Recognized by CrossParse

commander window query options	relevant string representation
implicit free sites	/F+
implicit ring closures	/R+
no isotopes	/N1
multifragment	/I2
charges	(#+<0,0,#-<0,0,#=<0,0)
radicals	#U<0,0
tautomers	/T1
separate fragments	/G1
absolute stereochemistry	/S1

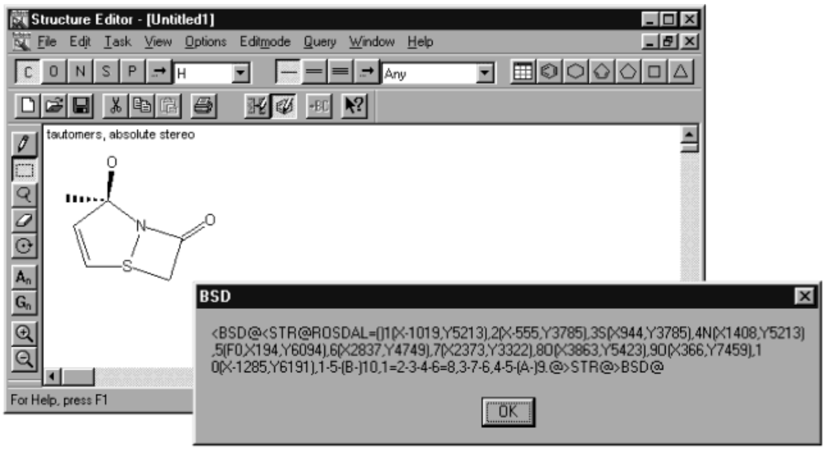
For those organizations using Gmelin by itself, or in conjunction with the Beilstein database, the same type of log can be produced and analyzed using CrossParse.

Each line in the transaction log shows a discrete activity by the end-user. These can be logging in to and out of the database, structure searching, reaction searching, fact searching, displaying of answers and using hyperlinks. The general format of the log file is represented in Table 3. The columns are separated by "|" (pipe symbol) in the log file. Each of the search statements found in the log file represents the full specification used by the end-user in his search, e.g. a structure search with atom lists, open free sites and stereochemistry.

Figure 1 shows a reaction search statement with highlighted specifications together with the reaction product molecule as a ROSDAL (the textual structure representation used by CrossFire) string. To view the ROSDAL string for a particular structure, when using the Beilstein structure editor, pressing <CTRL-ALT>B provides a pop-up box containing the ROSDAL representation. This is the ROSDAL string that will appear in the transaction log when the search is submitted (Figure 2).

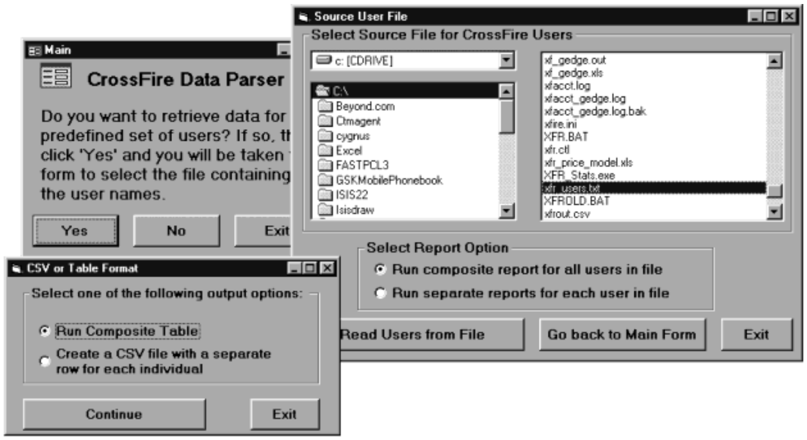
CrossParse can analyze for approximately 40 discrete specifications within a search statement. Many others can be derived using combinations of these discrete specifications, e.g. the use of a stereo bond in a structure without having turned on stereo searching via one of the Options menus (shown in Figure 5 as 'Stereo bond w/o attribute on').

Table 4 shows examples of some of the character strings that have specific meanings within the transaction log and hence can be searched using CrossParse to indicate when a user has drawn a structure with a particular feature present.



```
<BSD@<STR@ROSDAL=()1(X-1019,Y5213),2(X-555,Y3785),3S(X944,Y3785),4N(X1408,Y5213)
,5(F0,X194,Y6094),6(X2837,Y4749),7(X2373,Y3322),8O(X3863,Y5423),9O(X366,Y7459),1
0(X-1285,Y6191),1-5-(B-)10,1=2-3-4-6=8,3-7-6,4-5-(A-)9.@>STR@>BSD@
```

Figure 2. The Beilstein structure editor and the ROSDAL representation for the structure drawn. Beilstein Database: Copyright 1988−2002, Beilstein Institut zur Foerderung der Chemischen Wissenschaften Crossfire Software: Copyright 1995−2002, MDL Information Systems. All rights reserved. Materials used herein under permission.

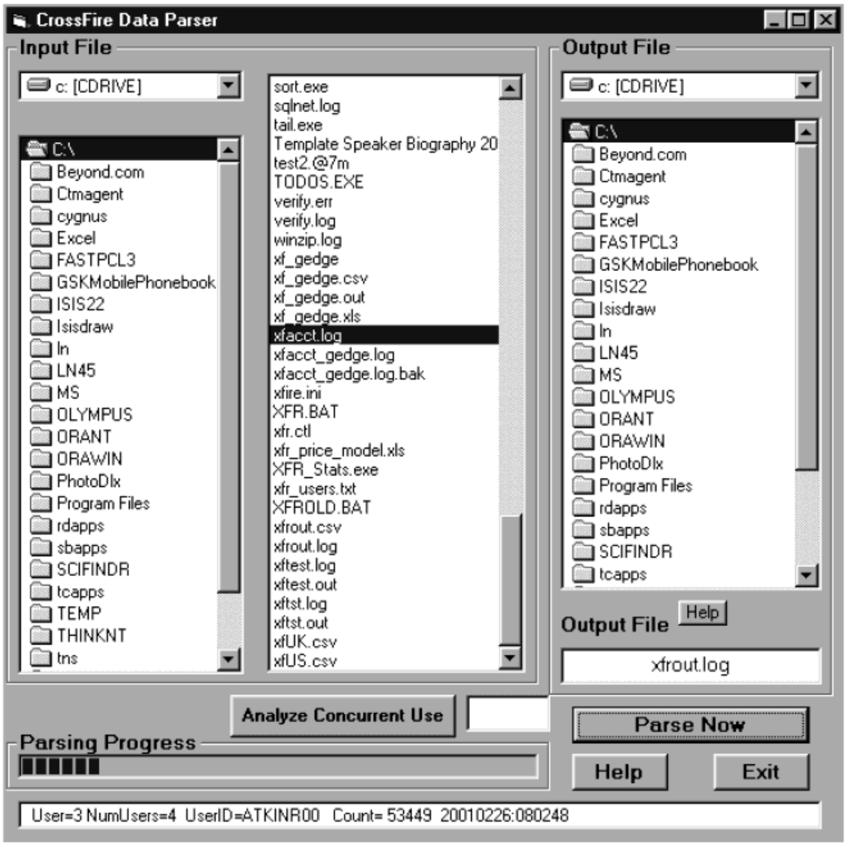


Do you want to retrieve data for predefined set of users? If so, th click 'Yes' and you will be taken form to select the file containing the user names.

Select Report Option
● Run composite report for all users in file
○ Run separate reports for each user in file

Select one of the following output options:
● Run Composite Table
○ Create a CSV file with a separate row for each individual

Figure 3. Setting up the requirements for a transaction log analysis. In this case a composite file is requested for all users in a specific group listed in the file xfr_users.txt.



Output File: xfrout.log

User=3 NumUsers=4 UserID=ATKINR00 Count= 53449 20010226:080248

Figure 4. Snapshot of CrossParse in use.

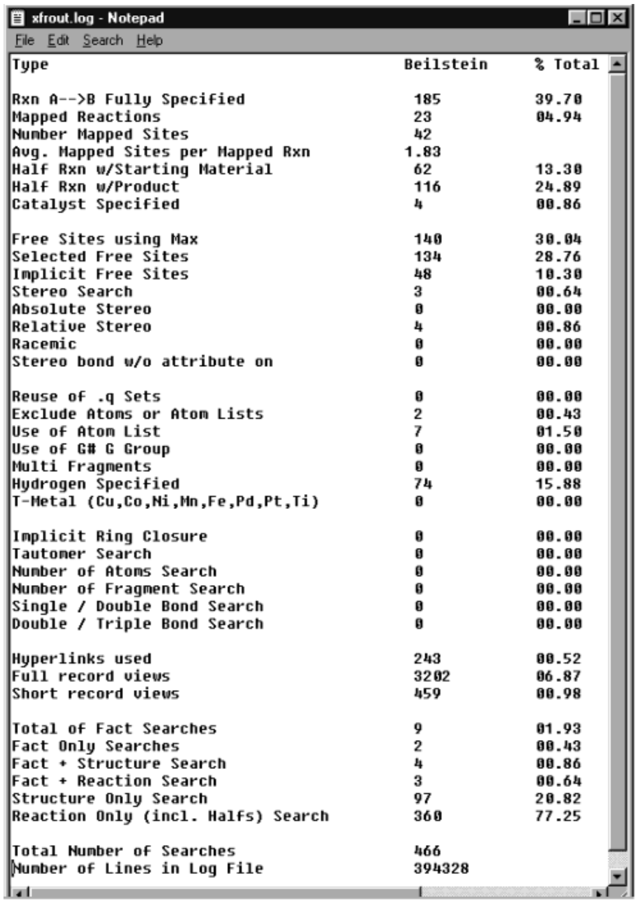
OPERATION OF CROSSPARSE

On starting the program the user is prompted to select a file of users, if so desired; this process is shown in Figure 3.

From here, the user can create a table of composite results for the analysis or create a CSV (comma separated values) file for each individual in the text file of users. A CSV file is readily opened in Microsoft Excel for data manipulation or graphing. The log file is then selected, the output file named and the 'Parse Now' button clicked.

Figure 4 shows CrossParse in use. The selection of an input file (the log file for analysis) and output file (where completed analysis data are recorded) is shown. The progress bar gives the user an indication of how long the process will take as some of these analyses can take many minutes or even hours for very large transaction logs. The bottom text box gives information about which user is being searched in a set of users and which line of the transaction log is currently being analyzed.

Typical output created from the log file analysis showing a list of various search specifications employed by the end-user(s) is given in Figure 5. Numbers of uses of specific search features or structure attributes are shown under the column headed "Beilstein", with the percentages of the total number of searches given in the right-hand (% Total) column. Note that the Hyperlink and Record view numbers in the % column refer to the average number of those activities per search.



```
Type                                   Beilstein     % Total

Rxn A-->B Fully Specified               185          39.70
Mapped Reactions                        23           04.94
Number Mapped Sites                     42
Avg. Mapped Sites per Mapped Rxn       1.83
Half Rxn w/Starting Material            62           13.30
Half Rxn w/Product                      116          24.89
Catalyst Specified                      4            00.86

Free Sites using Max                    140          30.04
Selected Free Sites                     134          28.76
Implicit Free Sites                     48           10.30
Stereo Search                           3            00.64
Absolute Stereo                         0            00.00
Relative Stereo                         4            00.86
Racemic                                 0            00.00
Stereo bond w/o attribute on            0            00.00

Reuse of .q Sets                        0            00.00
Exclude Atoms or Atom Lists             2            00.43
Use of Atom List                        7            01.50
Use of G# G Group                       0            00.00
Multi Fragments                         0            00.00
Hydrogen Specified                      74           15.88
T-Metal (Cu,Co,Ni,Mn,Fe,Pd,Pt,Ti)       0            00.00

Implicit Ring Closure                   0            00.00
Tautomer Search                         0            00.00
Number of Atoms Search                  0            00.00
Number of Fragment Search               0            00.00
Single / Double Bond Search             0            00.00
Double / Triple Bond Search             0            00.00

Hyperlinks used                         243          00.52
Full record views                       3202         06.87
Short record views                      459          00.98

Total of Fact Searches                  9            01.93
Fact Only Searches                      2            00.43
Fact + Structure Search                 4            00.86
Fact + Reaction Search                  3            00.64
Structure Only Search                   97           20.82
Reaction Only (incl. Halfs) Search      360          77.25

Total Number of Searches                466
Number of Lines in Log File             394328
```

Figure 5. Typical tabular output file from CrossParse.

USER STUDIES

Preliminary results from CrossParse have already been reported in the context of aiding database and application purchase decisions.[26] The present work incorporates a number of additional studies, a more in depth analysis and comparisons with traditional questionnaire surveys.

Study A − MIMAS. For this study, an investigation was made of the log files for the period Oct−Dec 2000. Three groups of users were identified. First, untrained users who had not attended a MIMAS CrossFire training course; this group comprised approximately 3000 users. Second, users who had been trained in 2000 but who had not subsequently made heavy use of CrossFire; this group contained a number of nonchemists and comprised a total of 70 users. Third, "Super users" who were users who had been trained in 2000 and who also make heavy use of CrossFire; this group size was 36. CrossParse was then run for the three groups of users. The results are shown in Figures 6−10.

Study B − SB. This study involved the parsing of two years worth of log files that represented approximately two million transactions logged by the server. At ex-SB, there were no regularly scheduled hands-on training events, hence, a comparison of the before and after searching behavior of the end-users was not possible. We chose to investigate the difference in searching behavior between the Synthetic and Medicinal chemists. The results are shown in Figures 6−10, alongside those from the MIMAS study A.

Study C − MIMAS: Questionnaire Results Compared to CrossParse Output. As described above, MIMAS in collaboration with GSK, devised two questionnaires, the first of which (Table 1) was distributed to attendees at three training courses in early 2001. Fifty-four questionnaires were returned. Only those where the participants had provided contact details (43) were followed up with the second questionnaire (Table 2) by e-mail. Nine e-mail questionnaires were completed, from which six were chosen for this study (three users had hardly used CrossFire prior to the training course and therefore were not included). The course questionnaires were examined in order to make comparisons between the users' perceived and actual search behavior.

Study D − MIMAS 2001 Course Attendees. An analysis of the usage of CrossFire made by all attendees at three MIMAS training courses in early 2001 was made. These users were the same as the "Trained" and "Super" users evaluated in Study A and totalled 106. CrossParse was used to analyze usage for three months prior to attending a course, three months immediately after the course, and a further two month period from July−August 2001. This study was undertaken to attempt to determine changes in usage patterns after a few months had elapsed since the training courses were attended, to see if training has a lasting effect or whether top-up or refresher courses might be beneficial. Results are presented in Figures 11 and 12.

CrossParse gives information about the number of fact searches done but does not list the types of fact searches performed. For this, a Unix script was written to look for all occurrences of fact searches and count them by type. This was done for the MIMAS population that was investigated in Study D. The results of this analysis show that as a percentage of the total number of searches run there were several types of fact searches whose usage increased markedly. These are indicated in Figure 13.

ANALYSIS OF RESULTS

Studies A and B. From the chart in Figure 6, it can be seen that the most common type of search undertaken by MIMAS CrossFire users is the structure search, accounting for 51% of the total number of searches made by untrained users. This figure reduces for the trained users, who appear to make greater use of the fact editor. This supports anecdotal evidence that indicates that the fact editor is the least used CrossFire facility: chemists have a natural affinity toward structure searches as structures and reactions are the natural language of chemistry and structure drawing is second nature to chemists. Reaction searches represent 35% of searches by untrained users, rising to 40% for "Super" users. This indicates that reaction searching may also be an area where a benefit is achieved from training; however, it should be remembered that the sample sizes of the trained and super users are small and therefore are not necessarily representative.

Figure 6 also shows that more than 50% of all searches by all ex-SB users are reaction searches and greater than 60% of all Synthetic Chemistry searches are reaction searches. Approximately 38% of all searches by all users are structure searches, and Medicinal Chemists are more likely to do structure searches than Synthetic Chemists. The dominant type of searching undertaken by ex-SB personnel is reaction searching, possibly indicating a better basic understanding of the overall functionality of the CrossFire system or, more likely, this is in line with their business need.

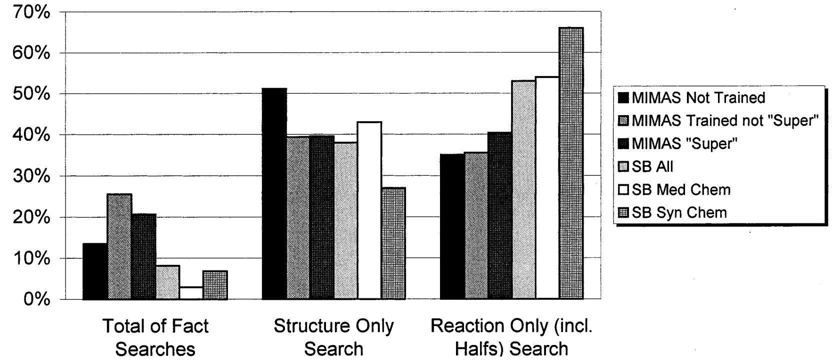


Figure 6. Results from CrossParse analysis for specific groups of MIMAS and GSK (ex-SB) users.

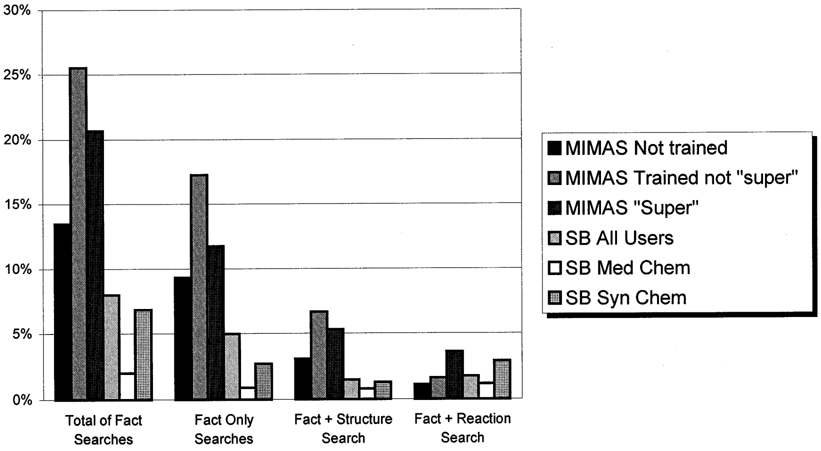


Figure 7. Output of CrossParse for CrossFire fact searching and combinations of fact searching with other types of search.

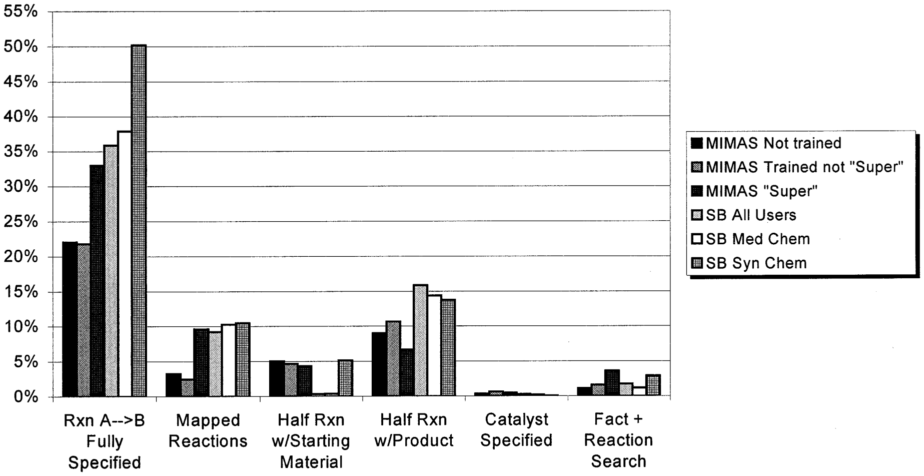


Figure 8. Output of CrossParse for reaction searching.

Figures 6 and 7 show that for GSK users, the use of the Fact Editor is lower than for the MIMAS users. This may be the result of the higher proportion of nonchemists in the MIMAS community, whose information requests necessitate bibliographic or property searches which are text based and therefore make use of the fact editor.

In ex-SB Fact Searches represent less than 10% of all searches, and Synthetic Chemists are more likely to do Fact searching than Medicinal Chemists. Fact only searches represents only 5% of all searches, and the primary users are the Information Management staff. Fact+Rxn and Fact+Structure searches represent a total of approximately 4% of all searches, and Synthetic Chemists are more likely to carry out Fact+Rxn searching. A breakdown of the nature of specific types of searches undertaken by the user groups in Studies A and B is given in Figures 7−10.

Figure 10 shows that stereo searching is used significantly more at MIMAS than ex-SB. This may reflect the purposeful message given to ex-SB scientists to avoid stereo searching unless absolutely necessary in order to avoid losing potentially valuable answers. Interestingly, many of the users of stereo features in the structure drawing program forget to switch on stereo searching in the Options menu hence no stereo searching is performed. This is particularly true for the few scientists at ex-SB who are trying to use this feature.

For ex-SB users, CrossParse indicates that substructure searching, using free sites set to maximum (Figure 9), accounts for 45% of all searches, and 50% of all Reaction

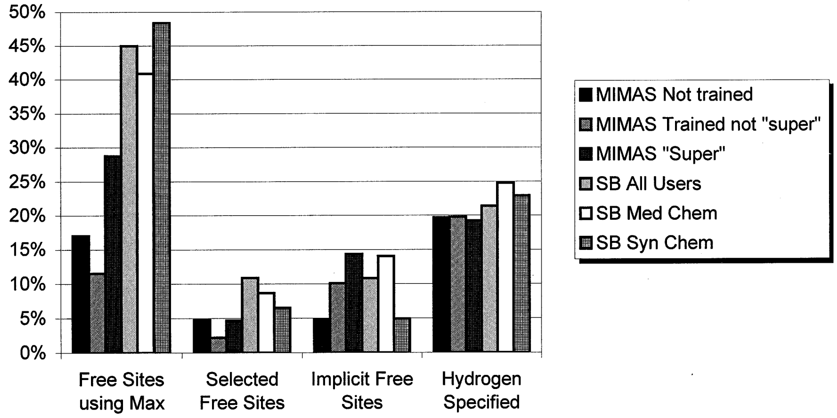


Figure 9. Output of CrossParse for a selection of structure building features. (*Free sites using max* = one or more specific sites within a structure have substitution allowed up to the maximum permitted by the valence; *Selected free sites* = one or more specific sites within a structure have substitution allowed up to a value specified by the user; *Implicit free sites* = full substructure search with substitution possible at all sites; *Hydrogen specified* = H atoms drawn rather than implicit.)

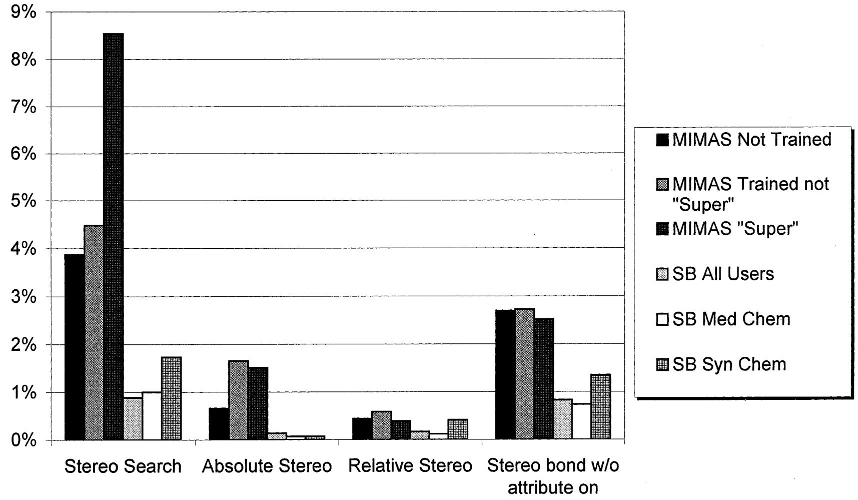


Figure 10. Output of CrossParse for stereo searching facility.

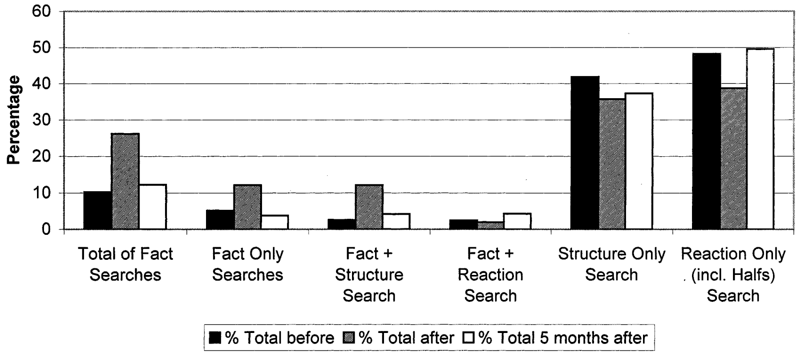


Figure 11. CrossFire usage as determined by CrossParse before, immediately after and five months after MIMAS CrossFire training courses.

Searches by Synthetic Chemists were A→B, with the corresponding percentage for Medicinal Chemists being 38% (Figures 8 and 9). Half reactions specifying products are more common than specifying starting materials, with ~10% and ~5%, respectively. This probably shows the emphasis on finding ways to make materials rather than finding ways in which a material might be used for other purposes. For example, a chemical company, having created a new chemical intermediate, might wish to look for outlets for this intermediate in other areas. Doing a substructure search using this intermediate as a starting material with no product specified could be very beneficial. MIMAS results indicate that training appears to increase the awareness of reaction mapping (Figure 8).

Study C. The results of CrossParse for the CrossFire usage made by six specific users who attended training courses in early 2001 were examined and compared with their answers to the before and after training questionnaires circulated (Tables 2 and 3).

CrossParse results showed an increase in the overall number of searches carried out of almost 60% over a similar time period. In addition, a number of functions were used

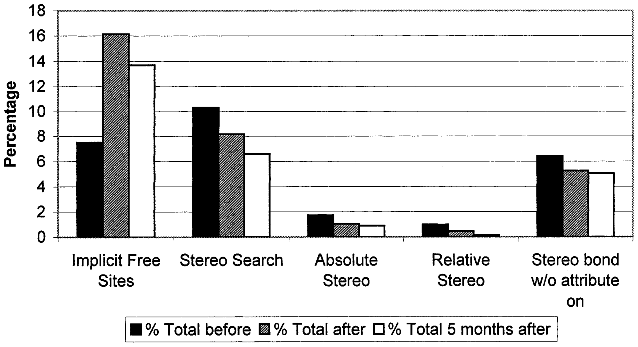


Figure 12. CrossFire structure searching behavior as determined by CrossParse before, immediately after and five months after the MIMAS CrossFire training course.

more after training than before. For example, 156 fact searches were done in the period after training, compared to 25 beforehand, and combinations of fact searches with reactions or structures also increased. The user-defined generic group functionality was used 16 times after the training course, whereas it was not used at all before training. The average number of mapped sites decreased from 4.45 per reaction (based on 49 reaction searches) to 2.91 (based on 93 reactions), suggesting a greater awareness of the role of reaction mapping and its usage. (It should be noted that there are many unmapped reactions in CrossFire Beilstein, and therefore routine mapping of reactions can be dangerous. Its use is recommended in cases where otherwise too many irrelevant hits are retrieved from a search.) Use of free sites, both implicit free sites to allow substitution at all sites in a compound and selected free sites to allow a partial substructure search, showed increases in usage. However, user-defined atom lists were not used either before or after the training course during the period examined. Some features showed little change in usage, e.g. stereochemistry, but it should be remembered that the sample of users was small and may not represent the whole user community so these features could simply not be needed by the group concerned.

The questionnaires indicate support of CrossParse evidence of greater use of CrossFire after the training sessions and also indicate that the trainees had tried out a number of the features identified by CrossParse as having been used more frequently. Most respondents to the questionnaires felt that their searching was also more efficient after the training course. This is an area that is difficult to analyze with CrossParse, as to some extent it is subjective.

Study D. It can be seen that several months after attending training courses, the usage of features which saw an initial rise has in some cases tended to revert to their previous levels, as shown in Figures 11 and 12.

Reasons for this change could be that the functions are not needed frequently—users tried them out immediately after the course and then have had little need for them since the training. Alternatively, the features could be difficult to use, and therefore if used infrequently the skills are lost. Although the sample here is relatively small, so caution must be applied when assessing these results, it seems reasonable to assume that periodically additional training might be desirable to reinforce the original learning from the course.

Figure 13 clearly shows a marked increase in the use of reagent (RX.RGT) searching as well as publication year (PY) and catalyst (RX.CAT). Interestingly, the use of chemical name (CN) decreased as did registry (RN) searching. This could be a direct effect of training which points out the potential deficiencies in searching these fields. However, the main impact of training appears to be an increased awareness of the range of fields that can be searched using the Fact editor. The total number of fields searched by the group prior to training is 17, immediately after training this rose to 36 and 4−5 months after training had reduced to 25. However, this represents a total of only 49 different fields searched by the group, out of a potential number of over 750.

LEARNINGS

From the above results, it appears that much of the functionality of CrossFire is underutilized. For example, less than 1% of all searches by all users involve specification of stereochemistry and less than 1% of all reaction searches specify catalysts. Other features which get little use include reuse of answer sets and some of the powerful structure query tools such as G groups and atom lists.

The Fact Editor is less frequently used than might be expected, bearing in mind the availability of the wealth of searchable property information on CrossFire which can only be used as search terms via the Fact Editor. (NB. Fact searches include all text-based searches, such as bibliographic details, numerical property searches, keyword searches of abstracts, etc.). However, the Commander 2000 Easy Data Search (EDS) forms might help to increase the use of this functionality. Additionally, few searches which combine fact and reaction or structure searches are undertaken. There are several potential reasons for this, namely that these features are so specialized that they are rarely needed, that users are unaware of them or that they are put off using them due to their apparent complexity.

Even though precision in reaction searching can be readily improved by the use of associated factual searches, e.g. for catalysts, this is not well used. (Users should, however, be made aware of the fact that reagent/catalyst/solvent information is not standardized in the text in CrossFire Beilstein, and use of the indexes is advised in order that all variations can be included.) When asked about this, a significant number of users prefer to leave the search broadly defined and to scroll through many answers. They are often unaware of how to improve precision by the use of factual parameters or are afraid of losing relevant answers by being overly restrictive in their search. Although fact searching may be low, much factual data may have been viewed. This may be more inefficient but is often the way a chemist would approach a problem, i.e., doing a structure search first and scrolling through a set of answers to find the appropriate hit with the right factual content.

Use of the Fact Editor by the MIMAS academic user community clearly increases after training, as shown in Figures 6, 7 and 13. Additionally, training leads to greater use of partial and full substructure searches (Figure 9). Anecdotal evidence suggests that some users are not aware that substructure searching is possible until this concept has been explained and demonstrated.

There is greater awareness of the diversity of factual searches that is possible through CrossFire, but this still represents only a small proportion of the potential fields which could be searched.

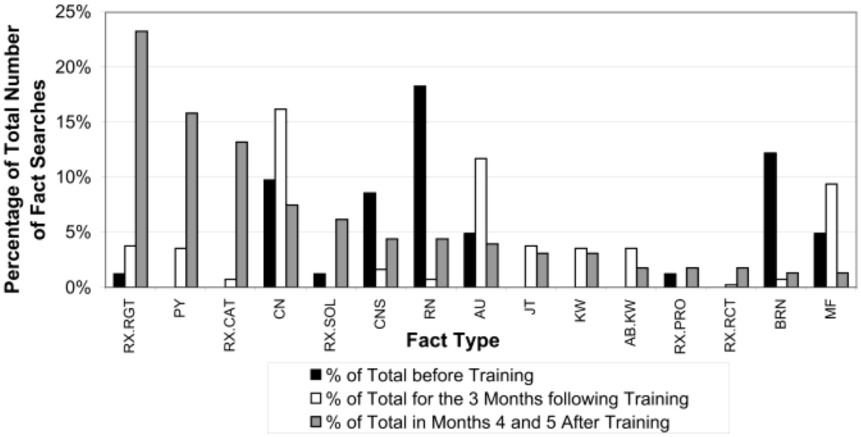


Figure 13. Results from the Unix script analysis of fact searching behavior of the Study D participants, showing the effect of time since training on fact searching behavior for 15 of the most heavily used search fields. (RX.RGT = reagent, PY = publication year, RX.CAT = catalyst, CN = chemical name, RX.SOL = solvent, CNS = chemical name segment, RN = CAS registry number, AU = author, JT = journal title, KW = keyword, AB.KW = abstract keyword, RX.PRO = product, RX.RCT = reactant, BRN = Beilstein registry number, MF = molecular formula.)

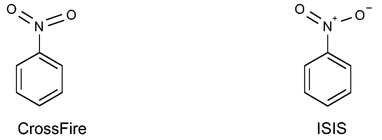


Figure 14. CrossFire and ISIS nitro conventions.

The analysis allows identification of user misconceptions. For example, it appears that implicit hydrogen atoms are specified in structures more often than would be expected (about 20% of all searches had hydrogen atoms specified), see Figure 9. This is likely to be a training issue, as is mapping of reactions (Figure 8), as fewer reactions appear to be mapped than would be expected. It is of some concern that many compounds searched for with stereochemistry specified in the structure drawing are done without having turned on the option to do this.

CrossParse has the ability to pull out specific structure searches that have potentially incorrect structure conventions embedded in them. One classic example is the nitro group for which the differing CrossFire and ISIS convention are shown in Figure 14. Given that a number of chemists at ex-SB use ISIS/Draw as their preferred structure editor, there is the potential to search on the charge-separated ISIS nitro, leading to erroneous results. CrossParse can write out to a file those structures that are suspect. Visual inspection can confirm or deny this. Even though this convention problem is covered in our training, some examples of the use of the wrong convention are seen in the logs. CrossParse could be extended to look for other structures with improper conventions applied e.g. sulfone, sulfoxide, azide, nitrogen oxides and phosphorus oxides.

Some features are rarely used, even after training, e.g. generic groups, atom lists, reuse of answer sets. The results for some of these functions have not been included in the charts as the percentages are close to zero. This could be due to the infrequent need for these functions or their complexity of use. Since these are complex requirements of the system, it is difficult to envisage how their use could be significantly simplified.

From Study D, it would appear that follow-up training courses could be beneficial to reinforce some of the skills learned during the initial training. In view of the difficulty in providing such training to a changing community of users at around 85 sites in the UK and overseas, it has been decided to provide this in the form of distance learning materials. MIMAS has therefore developed Web-based self-paced tutorials for the UK, Irish and Scandinavian higher education communities which are available at all times to aid CrossFire users whenever assistance is required.

ANALYSIS OF CONCURRENT USE

Another benefit of CrossParse is the ability to generate concurrent user statistics from the raw data. Typical concurrent user statistics generated from a server simply give the number of users logged into the application at any given time. However, some users often open the CrossFire application early in the day and stay logged in until timed out with very little activity. (The time out period is set by the CrossFire administrator. The period might vary between several minutes and several hours.) A better measure of true concurrent use, at least where server capability is concerned, is the number of ‘active’ users of the system.

By analyzing the transaction log file for active users and the particular times they are using the system we can generate a time versus concurrent user map. CrossParse works by measuring one minute time slices which have some user activity and measuring unique users per time slice. Only time slices with activity are reported, hence, the chart in Figure 15 shows a minimum activity of 1 and not 0.

This approach has proved very valuable during the merger to form GSK. Several CrossFire servers were monitored for their concurrent use, adjusted for time differences and concurrent usage added together. This gave us a target for a GSK single server deployment. With appropriate high level concurrent testing on our highest performance server we were able to demonstrate that it could take the total GSK load. Hence, the other servers were retired with considerable associated savings. Also, the concurrent use analysis provides useful insight into peak times of activity during the day and when regular maintenance might best be done i.e., during low load times. Other studies have investigated concurrent use in order to better spread the load on their application server.[27]

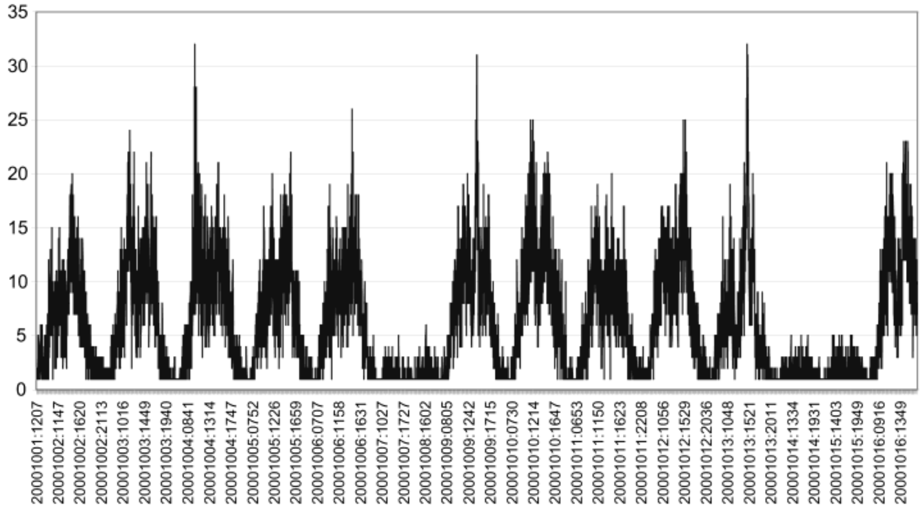


Figure 15. Analysis of MIMAS concurrent use of CrossFire, showing the number of unique users per one minute time slice.

CONCLUSIONS

Internal. The results of the CrossParse analysis will be used in further training development, including targeting courses to specific types of query and user. CrossParse, therefore, helps determine learning objectives of training courses for specific audiences.

It is clear that it is easy for some users to have misconceptions about structure and reaction drawing facilities and how searches are executed, e.g. implicit hydrogens, stereochemical searching, barriers to use of Fact Searching. CrossParse has highlighted these, and therefore these issues can be given extra emphasis in training sessions.

Misconceptions can sometimes arise due to users having experience of a system that employs different conventions, and as these can be identified using CrossParse, there are clearly benefits for cross-application training.

Indications of where support documentation needs modification have also been obtained.

In the case of the multiple GSK CrossFire servers, CrossParse gave crucial information for the proposal to consolidate servers. This tool allowed calculated resource optimization at minimal risk.

For MDL. It would seem from the results of the CrossParse analysis that some aspects of CrossFire may benefit from being made simpler, e.g. generic group searching, atom lists, catalyst specification, manipulation of answer sets. Additionally, some defaults might not be correct. For example, should the default for a structure search be an exact match or a substructure search? Should the default when a stereo bond is drawn be “stereo off” or “stereo on”?

For Other Vendors. It would be helpful if other vendors were to provide utilities such as this VB application in order that all customers could benefit from the extra knowledge obtainable from the log files generated by these applications. If this is not possible, then at least the option to generate detailed transaction logs should be made possible in order that organizations can produce their own parsing programs.

Future Work. The CrossFire logs are a rich source of user navigation habits. Some work has begun to explore whether these navigation patterns can be utilized in further design enhancements to the application as well as an aid in training development. A mathematical/statistical approach to this has been undertaken by Spiliopoulou.[28]

Additionally, with the release of CrossFire 2000 and its enhanced fact searching form, there is an opportunity to investigate whether this has had a noticeable effect on fact searching by our end-users. Also, with new releases of the CrossFire client, CrossParse will give us the opportunity to more fully evaluate the usefulness of any new features by comparing before and after behavior patterns.

REFERENCES AND NOTES

(1) McGlamery, P. MAGIC transaction logs as measures of access, use and community. *J. Acad. Librarianship*. **1997**, *23*, 505−510.

(2) Atlas, M.; Little, K. R.; Purcell, M. O. Flip charts at the OPAC: using transaction log analysis to judge their effectiveness. *Ref. User Serv. Q.* **1997**, *37*, 63−69.

(3) Jocic, M. Analysis of users’ searches of CD-ROM databases in the national and university library in Zagreb. *Inf. Process. Manage*. **1997**, *33*, 785−802.

(4) Millsap, L.; Ferl, T. E. Search patterns of remote users: an analysis of OPAC transaction logs. *Inf. Technol. Libr.* **1993**, *12*, 321−343.

(5) Bangalore, N. S. Re-engineering the OPAC using transaction logs. *Libri* **1997**, *47*, 67−76.

(6) Blecic, D. D.; Bangalore, N. S.; Dorsch, J. L.; Henderson, C. L.; Koenig, M. H.; Weller, A. C. Using transaction log analysis to improve OPAC retrieval results. *College Res. Libr.* **1998**, *59*, 39−50.

(7) Bishop, A. P.; Neumann, L. J.; Star, S. L.; Merkel, C.; Ignacio, E.; Sandusky, R. J. Digital libraries: Situating use in changing information infrastructure. *J. Am. Soc. Inf. Sci.* **2000**, *51*, 394−413.

(8) Borgman, C. L.; Hirsh, S. G.; Hiller, J. Rethinking online monitoring methods for information retrieval systems − from search product to search process. *J. Am. Soc. Inf. Sci.* **1996**, *47*, 568−583.

(9) Ciliberti, A.; Radford, M. L.; Radford, G. P.; Ballard, T. Empty handed − a material availability study and transaction log analysis verification. *J. Acad. Librarianship.* **1998**, *24*, 282−289.

(10) Connaway, L. S.; Budd, J. M.; Kochtanek, T. R. An investigation of the use of an online catalog − user characteristics and transaction log analysis. *Libr. Resources Technol. Serv.* **1995**, *39*, 143−152.

(11) Ferl, T. E.; Millsap, L. The knuckle-crackers dilemma − a transaction log study of opac subject searching. *Inf. Technol. Libr.* **1996**, *15*, 81−98.

(12) King, N. S. End-user errors − a content analysis of paperchase transaction logs. *Bull. Med. Libr. Assoc.* **1993**, *81*, 439−441.

(13) Nicholas, D. An assessment of the online searching behavior of practitioner end users. *J. Doc.* **1996**, *52*, 227−251.

(14) Wallace, P. M. How do patrons search the online catalog when no ones looking − transaction log analysis and implications for bibliographic instruction and system design. *RQ*. **1993**, *33*, 239−252.

(15) Moukdad, H.; Large, A. Users’ perceptions of the Web as revealed by transaction log analysis. *Online Inf. Rev.* **2001**, *25*, 349−358.

(16) Wyly, B. J. From access points to materials: a transaction log analysis of access point value for online catalog users. *Libr. Resources Technol. Serv.* **1996**, *40*, 211−236.

(17) Jones, S.; Gatford, M.; Do, T.; Walker, S. Transaction logging. *J. Doc.* **1997**, *53*, 35−50.

(18) Reiter, M. B. Can you teach me to do my own searching? Or tailoring online training to the needs of the end-user. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 419−422.
(19) Ostrum, G. K. and Yoder, D. K. Chemists as end-user searchers − training and follow up. Online Information: 9th International Meeting, 1985, pp 131−140.
(20) Oldroyd, B. K. Study strategies used in online searching 5: differences between the experienced and inexperienced searcher. *Online Rev.* **1984**, *8*, 233−244.
(21) Mullan, N. A.; Blick, A. R. Initial experiences of untrained end-users with a life-sciences CD-ROM database: a salutary experience. *J. Inf. Sci.* **1987**, *13*, 139−141.
(22) MDL Information Systems GmbH. www.Beilstein.com.
(23) MIMAS. www.mimas.ac.uk.
(24) Meehan, P.; Schofield, H. CrossFire: a structural revolution for chemists. *Online Inf. Rev.* **2001**, *25*, 241−249.
(25) JISC, a strategic advisory committee working on behalf of the funding bodies for further and higher education in England, Scotland, Wales and Northern Ireland. www.jisc.ac.uk.
(26) Cooke, F.; Schofield, H. Mining for information nuggets: assessment of end-users' search techniques to assist with training and purchase decisions. Proceedings of the 2001 International Chemical Information Conference, Nîmes, France, 21−24 October 2001, pp 7−16.
(27) Lucas, T. A. Time patterns in remote OPAC use. *College Res. Libr*. September **1993**, *54*, 439−445.
(28) Spiliopoulou, M. The laborious way from data mining to web log mining. *Comput. Syst. Sci. Eng.* **1999**, *14*, 113−126.

CI0202638