

# **E-Journal User Study**

## **Report of Web Log Data Mining**

### **December 2002**

#### **1. INTRODUCTION**

The emergence of e-journals has provided new methods and tools for scientific research—not only for scientists seeking articles, but also for publishers and others trying to understand subscribers’ behaviors. E-journals generate log files that track the steps of readers as they move around on journal websites. Analyzing these web log files is very challenging, however, due to their huge size (typically millions of lines of data) and the heterogeneity of the usage patterns of each individual. Finding useful information buried in such large-scale datasets requires a rigorous analytical approach, yet the data are not very amenable to classical statistical models.

Data mining is the term currently used for analyses which explore possible patterns or behaviors in datasets without utilizing classical hypothesis-driven statistical models (Hand et. al., 2001). Though it is very difficult to identify representative patterns of users’ behaviors on the Web, previous studies have attempted to develop rigorous analytical tools to investigate web visit patterns. Huberman et. al. (1998) demonstrated that Web visits have some unique and regular patterns, particularly those related to the number of clicks (hyperlink requests). Cadez et. al. (2000) explored a tool to describe such patterns in a visual way and defined statistical models for use in analyzing the data.

Unlike other research about the behaviors of individuals visiting many kinds of websites, this analysis focused on visits to scientific e-journal websites. Such usage usually has (1) a very specific purpose—retrieving scientific information such as articles and databases, and (2) relatively homogenous user populations which are likely to be similar to one another in their reasons for using e-journals.<sup>1</sup> As a result, we can identify typical usage patterns relatively easily. Our analysis revealed two patterns of e-journal website visits: downloading/retrieving full text and accessing other online features such as databases or other research resources.

This part of the E-Journal User Study (eJUSt) analyzes web log data from 14 medical and life sciences journals to gain insight about e-journal users’ behaviors on the Web. This report uses basic analytic tools to help understand users’ website behaviors by (1) identifying typical usage patterns; and (2) analyzing specific examples in a “detailed” way. We summarize the data in

---

<sup>1</sup> Though life scientists and clinicians differ from each other in needs and uses for e-journals (as discussed in other reports of this project, especially the first survey report), they are relatively homogeneous compared to the general populations examined in most Web usage studies, which may include shoppers, schoolchildren, criminals, patients, rocket scientists (literal and otherwise) along with academics and clinicians).

Section 2 below, present the findings in Section 3, draw conclusions in Section 4, and discuss possible implications in Section 5.

## 2. DATA SUMMARY

### I. Data collection period

The data collection period was February 13, 2002. We chose this date after analyzing data about traffic patterns to see whether they varied with each journal's issue circulation<sup>2</sup> period. There was a clear and significant weekly cycle in terms of traffic (number of sessions) regardless of the length of issue circulation period. Monday and Tuesday had the highest traffic and then traffic slowed down until Friday. So we picked a day in the middle of the week, Wednesday, to collect our sample for analysis. One-day sampling is sufficient enough to represent the targeted population in investigating usage patterns and users' behaviors on the Web. Hand et. al. discuss the need to reduce the massive size of log file datasets to a more manageable size for analysis, and note that data mining algorithms operate on an approximate version of the full data set. For the approximate version of our dataset, we chose to analyze web log data from February 13, 2002, a moderate-traffic day, for all 14 journals, generating a manageable dataset for data mining.

### II. Examples of data

Here are a few examples from actual lines of web log data. The log data contain variables such as IP address, request date and time, response code, web address referred by, bytes transferred, and request page.

```
132.236.171.212,02/18/02,12:46:18,300,http://www.journal.org,8372,/search.dtl
```

```
132.236.171.212,02/18/02,12:46:25,300,http://www.journal.org/search.dtl,40557  
/,/cgi/search?volume=&firstpage=&author1=pueschel&author2=&titleabstract=&full  
text=&fmonth=Feb&fyear=1998&tmonth=Feb&tyear=2002&hits=10&sendit=Search&journ  
alcode=journal&fdatedef=1+February+1998&tdatedef=1+February+2002
```

```
132.239.1.232,02/15/02,07:48:22,200,-,9869,/cgi/content/abstract/38/1/107
```

```
132.239.1.232,02/15/02,07:49:47,200,-,9834,/cgi/content/abstract/38/1/164
```

```
132.239.237.38,03/14/02,05:27:07,200,http://www.google.com/search?hl=en&ie=IS  
O-8859-1&oe=ISO-8859-1&q=Journal+of+journal name,7000,/  

```

```
132.239.237.38,03/14/02,05:27:13,200,http://www.journal.org,6343,/contents-  
by-date.0.shtml
```

### III. Participating journals

---

<sup>2</sup> Journal issue circulation period is defined as the period between when a new issue becomes available on the website and when the next issue becomes available online.

For our analysis, we used web log data from 14 journals in the life sciences and medicine associated with HighWire Press. Table 1 lists the 14 participating journals and gives the average number of sessions per day and the issue circulation period of each journal.

**Table 1: Participating Journals: summary of web log data on February 13, 2002**

<b>Journal Name</b>	<b>Sessions per day--weekday (mean)</b>	<b>Issue circulation period (Feb. 2002)</b>
<b>British Journal of Psychiatry</b>	2,000	1/31-/27
<b>Blood</b>	6,026	2/5-2/19
<b>Chest</b>	2,887	2/7-3/10
<b>Circulation</b>	6,520	2/11-2/18
<b>Clinical Chemistry</b>	1,313	1/22-2/27
<b>Endocrinology</b>	1,783	1/16-2/20
<b>Journal of Applied Physiology</b>	1,985	2/12-3/14
<b>Journal of Biological Chemistry</b>	21,855	2/8-2/14
<b>Journal of Clinical Microbiology</b>	1,188	2/1-2/28
<b>Journal of Immunology</b>	4,328	1/31-2/19
<b>Journal of Nutrition</b>	2,972	1/31-3/14
<b>Journal of Pharmacology &amp; ET</b>	1,820	1/18-2/20
<b>Plant Cell</b>	2,167	2/1-3/6
<b>Radiology</b>	1,408	1/29-2/25

#### **IV. Session Definition**

As is clear from examples of data in section 2.II, the original log data show sequences of requests by IP address and request time. We are interested in individual behaviors, such as how an individual initiates his/her web visits, how many pages this individual requests while he/she is on the web, which web pages he/she requests, and the pattern of requests.

To analyze these individual behaviors, we need to first define an individual session. A session can be defined by a unique IP address and a unique request time. When we first come across an IP address, we note the request time and define that as the beginning of a session. We then keep tracking that individual's requests continuously, and we define the end of that particular session for that individual to be when a subsequent request does not appear within an hour. Thus a particular individual could visit the same website multiple times during the day, but if intervals between these visits are longer than an hour, then we consider these different sessions.

One limitation to this approach is that when a single computer is shared by many users, one person may visit a journal's website and finish his session but leave the computer on with the same web page open. Then a different person could show up (within an hour) and surf the same web page which is already open. Under our definition, this would be recorded as a single session even though it contains the patterns of two different individuals. An example of this case is a laboratory sharing a few computers for online access among all lab members. However, defining a session is a necessary step for analysis to investigate usage patterns, and we assume that this occurrence is relatively rare, especially within our hour-long session timeframe.

## V. Category Definition

The Web log data we collected are relatively homogenous, since the websites were all developed in conjunction with HighWire Press and the web log data were also collected by HighWire. We found six major request types in the journal web visits: Journal Home Page, Table of Contents, Full Text in HTML version, Full Text in PDF format, Search, and selected online Hyperlinking features (including hyperlinks to cited articles, to articles in press, and to field-specific resources/databases). These six categories were used to frame our analysis.

## VI. Traffic

This section describes the average traffic volume per session. Traffic volume differed somewhat by journal. The average and median numbers of requests per session ranged from 5 to 8 and from 4 to 5, respectively (Table 2). That average request numbers are much higher than the medians indicates that the distribution of numbers of requests was right-skewed, i.e., users who make more requests on the Web make many more requests than users who make fewer requests.

**Table 2: Session Traffic by Journal**

Journals(# of sessions per day)	Mean (Median)	Inter-quartile (25%-75%)	Only One Click Per Session
<b>British Journal of Psychiatry (1,720)</b>	5 (2)	1-6	36%
<b>Blood (6,026)</b>	7 (4)	1-8	19%
<b>Chest (2,887)</b>	7 (4)	2-8	22%
<b>Circulation (6,520)</b>	8 (4)	2-9	21%
<b>Clinical Chemistry (1,313)</b>	6 (4)	1-7	25%
<b>Endocrinology (1,783)</b>	5 (4)	2-6	20%
<b>Journal of Biological Chemistry (21,855)</b>	8 (5)	3-8	10%
<b>Journal of Clinical Microbiology (1188)</b>	7 (4)	2-7	24%
<b>Journal of Applied Physiology (1,985)</b>	6 (4)	2-7	24%
<b>Journal of Immunology (4,328)</b>	7(4)	2-8	10%
<b>Journal of Pharmacology E&amp;T (1,820)</b>	5.5(4)	2-6	19%

<b>Journal of Nutrition (2,972)</b>	5.5(3)	1-6	31%
<b>Plant Cell (2,167)</b>	5(3)	1-6	32%
<b>Radiology (1,408)</b>	8(4)	2-8	23%

A significant percentage of sessions (3<sup>rd</sup> column, Table 2) lasted for only one request (10% to 36%, varying by journal).

Table 3 shows the average duration of each request by journal—how long each request lasts. We did not measure the duration of the last request of each session, since we do not know how long the reader continued with that last page (that is, we cannot know from web log data at what point the session actually ends or how), and we considered only sessions with more than one request, for the same reason. The average duration of each request varied greatly by journal, with a range of 5 minutes to 25 minutes.

**Table 3: Elapsed Seconds per request**

<b>Journals (# of sessions per day)</b>	<b>Mean ± Standard Deviation</b>	<b>Median</b>	<b>Min</b>	<b>Max</b>
<b>British Journal of Psychology (1,720)</b>	314 ± 177	352	1	542
<b>Blood (6,026)</b>	922 ± 533	1010	1	1628
<b>Chest (2,887)</b>	532 ± 320	573	1	965
<b>Circulation (6,520)</b>	1072 ± 679	1141	1	1986
<b>Clinical Chemistry (1,313)</b>	304 ± 184	333	1	546
<b>Endocrinology (7,615)</b>	354 ± 234	389	1	647
<b>Journal of Biological Chemistry (21,855)</b>	1350 ± 898	1419	1	2627
<b>Journal of Clinical Microbiology (1,938)</b>	462 ± 286	481	1	857
<b>Journal of Applied Physiology (1,985)</b>	426 ± 249	471	1	749
<b>Journal of Immunology (4,328)</b>	857 ± 499	966	1	1504
<b>Journal of Pharmacology E&amp;T (1,820)</b>	391 ± 243	438	1	690
<b>Journal of Nutrition (2,972)</b>	406 ± 311	400	1	838
<b>Plant Cell (2,167)</b>	480 ± 293	536	1	867
<b>Radiology (1,408)</b>	356 ± 233	361	1	691

Note: The last click per session was not included in the calculation of statistics; only sessions that lasted more than one request were included.

## **VII. Demographic characteristics: country of residence, institutions, unidentified IP addresses**

IP addresses can provide information on the countries and institutions from which requests originate. Each journal had a different pattern of users' locations, but roughly 30-40% of requests were from outside the United States (for journals based in the United States). The data on users'

institutions showed that medical journals appear to attract users from pharmaceutical companies more than biology journals do.

*Journal of Nutrition* had the smallest percentage of foreign users while *Endocrinology*, *Chest* and the *Journal of Immunology* had relatively more requests originating from foreign countries.

The percentage of users from U.S. academic institutions also varied by journal: 23% for *Journal of Biological Chemistry*; 23% for *Journal of Immunology*; 25% for *Journal of Applied Physiology*; 22% for *Clinical Chemistry* and *Journal of Clinical Microbiology*; and 8% for *British Journal of Psychiatry* (refer Table A.3. in Appendix).

Because many IP addresses were not identifiable (68-80 % of sessions had identified IP addresses), all these percentages must be analyzed with caution. Rates of unidentifiable IP addresses ranged from 20% (*British Journal of Psychiatry*) to 32% (*Radiology*) of sessions.

### **VIII. Referrals**

Web log data contain information about how each session was referred to a journal website (“web address referred by” in Section 2.II.). The referrals in our data set can be categorized as by institutional websites (such as educational institutions and companies), by generic search engines such as google.com, and by society websites. We found that library websites listing available online journals were responsible for a significant amount of journal traffic (10-25% of sessions, depending on the journal), so libraries appear to play an important role in attracting traffic to e-journal websites. Such referral addresses were identifiable in such forms as “*ejournal.library.XXX.edu*”, “*catalog.library.XXX.edu*” or “*digitallibrary.XXX.edu*”

## **3. FINDINGS**

### **I. Users’ Sequences**

Before we discuss usage patterns, we need to discuss our findings on the starting and ending points of sessions. We will then use these data with a transition matrix to define a full sequence of usage.

#### **A. Starting Points of Usage Sequences**

*The two most common ways for users to visit journal websites for online full-text retrieval were visiting the journal home page directly, and being directed by PubMed.*

There were two major starting points for journal web visits—through journal home pages and through PubMed. Some journals’ traffic came mostly from home pages, while *Journal of Biological Chemistry* had major traffic from PubMed. Different starting points lead to different user behaviors, because the choices of subsequent page requests are limited by the current page.

*Entering journal websites through homepages usually leads to either browsing contents or searching for an article.*

A session starting at a journal homepage has only a few possibilities for the next request: (1) browsing tables of contents; (2) searching for an article; or (3) hyperlinking to online journal features. We found that moving from a journal homepage to choice 1 or choice 2 (browsing contents and searching for articles) were the most frequent patterns of initial usage sequences.

*Users tend to read full text after browsing contents.*

More users read full text right away instead of reading abstracts first to see if articles are of interest; however, certain journals' users requested abstracts before reading full text (presumably to find out whether articles are of interest) more than other journals' users. Web visitors to the *British Journal of Psychiatry*, the *Journal of Applied Physiology*, and the *Journal of Clinical Microbiology* requested abstracts in 11% to 20 % of total clicks while other journals requested abstracts in only 3% to 7% of total clicks.

*More than 95% of sessions referred through PubMed request full text in HTML.*

PubMed is the other major source of traffic to journal websites. Most sessions directed by PubMed go to full text in HTML directly, because PubMed links direct sessions to HTML versions of full-text articles by default. As PubMed provides abstracts for most or all articles in the journals studied, the goal of users who visit journal websites through PubMed is very likely to be retrieval and download of full-text articles.

*Individuals often request PDF for printing/archiving after reading full text in HTML.*

Users seem to prefer PDF for article retrieval/download. They often first read HTML versions of full-text articles and then request PDF versions. This suggests that users may use PDF versions for printing or for archiving articles to a file in their computers. PDF appears to be a "final destination" for many sessions.

*Either abstracts or full text in HTML precede requests for full text in PDF format.*

It is clear that requesting PDF versions of full-text articles is a final destination of web visits after finding articles of interest. Users may reach this destination via either the abstract or the full text in HTML form. .

*Visits to e-journals through PubMed typically end shortly after full-text retrieval.*

This is not a surprising result, considering the typical purpose of web visits through PubMed. PubMed visitors likely have one goal in common: to retrieve/download full-text articles from the Web. After a successful retrieval, they leave the journal's website to continue their searches on PubMed's website. On average, the number of requests redirected by PubMed is 2 to 8, while other sessions, such as sessions which start from journal homepages, requested on average 4 to 10 clicks.

## B. Ending Points

*The final goal of e-journal website visits is most likely full-text printing.*

Analysis of usage patterns shows that the ending points (final requests) of web sessions are most likely to be requests for PDF versions of full-text articles.

All three of the most common usage patterns (please refer to Figures 1, 2, and 3 below) ended with requests for PDF, regardless of the starting or middle points of the sequences. PDF format is particularly well suited for printing journal articles, strongly suggesting that the final goal of journal web visits is article printing. 68% of our follow-up survey respondents said that they immediately print out a full-text article upon retrieving it online and read the printed copy.<sup>3</sup>

*A primary goal of e-journal website visits is to retrieve full text online.*

More than 60% of sessions finish by requesting full text (either in HTML or PDF). There is little variation by journal in which format is requested. The journal *Circulation* and *Radiology* has somewhat more sessions ending with HTML than PDF, while sessions from most other journals are somewhat more likely to end with PDF.

Compared to biology journals, medical journals seem more likely to have more traffic through journal homepages than through PubMed; however, this can not be generalized with much certainty in this small sample of journals. Some biology journals, such as *Journal of Nutrition*, have more traffic through their homepage than through PubMed.

## C. Typical Usage Patterns

The next step in analysis is selecting a suitable statistical model to describe usage patterns. We used a Markov Chain (MC) model where we model the probability that a user will go to a certain page given he/she is viewing the current page. For example, we model  $P(\text{Table of Contents} | \text{Home Page}) = \text{probability of requesting "Table of Contents" given the current page is "home page"}$ , and  $P(\text{PDF} | \text{HTML}) = \text{probability of requesting "PDF" given the current page is "HTML"}$ , etc. To model sequences, we have a **transition matrix** of size  $d*d$  (where  $d$  is the number of categories), and this varies by journal.

We selected several categories for formal analysis (see section 3.II for definition of categories).<sup>4</sup> Tables A.1 and A.2 in the Appendix show the transition matrix for each journal. (Rows of the matrix do not add up to 100% because we omitted categories that we initially included in the analysis but did not find to be statistically significant enough to discuss, having less than 5% of clicks out of total clicks.) Here are some examples of common usage sequences which stood out significantly compared to the rest of the sequences:

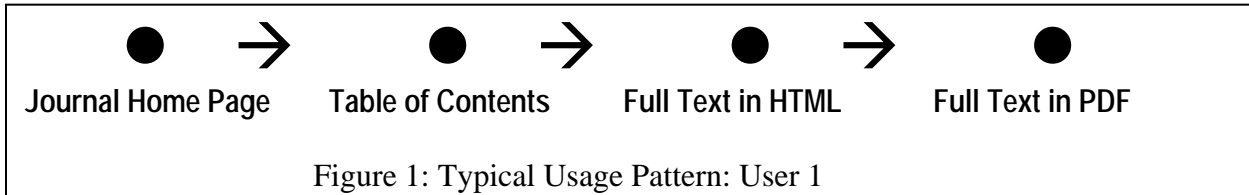
---

<sup>3</sup> See eJUSt follow-up survey report pp. 15; [http://ejust.stanford.edu/findings3/report\\_survey3.pdf](http://ejust.stanford.edu/findings3/report_survey3.pdf)

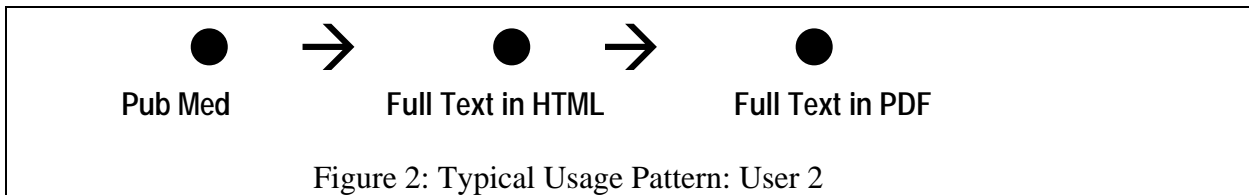
<sup>4</sup> We ran a sensitivity test and selected down to a few categories with significant numbers of usage. First we ran it with all categories for the transition matrix, selecting major categories by usage, then we repeated the same procedure with a few selected categories having significant usage proportions.

- Homepage visits leading to either searching for articles or browsing contents.
- Searches followed by another search then eventually leading to full text.
- Requesting an abstract followed by another abstract; then leading to PDF.
- Full text in HTML followed by requesting full text in PDF.

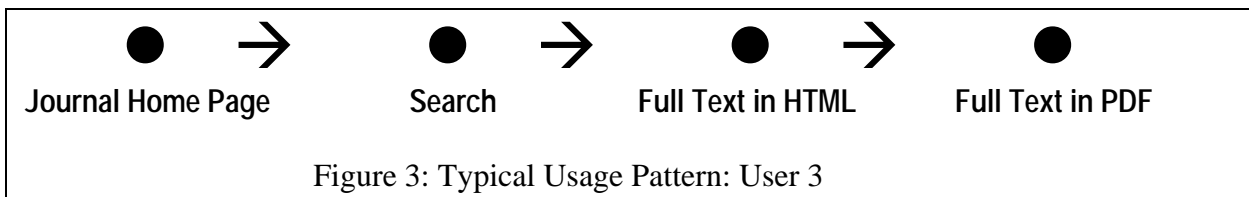
We found many different patterns in usage, but three typical patterns summarize the sequences well. The next three figures describe these typical usage sequences. User 1 in Figure 1 below comes to the journal website through the journal’s home page and browses either the current issue or back issues on the Web to find articles of interest, then prints them out.



Users who enter through Pub Med (User 2 in Figure 2 below) leave the website shortly after retrieving full text. PubMed directs users to full text in HTML and then users print out full text in PDF format. Users leave the website shortly after full-text retrieval.



Another typical usage pattern starting through journal homepages involves searching for an article within a journal. User 3 in Figure 3 prefers to locate articles using the “search” function. This might be either because search helps locate specific articles faster than browsing contents, or because the user has limited information about the article(s) sought but knows the typical material covered by the journal. Searches on a journal website are most helpful if users have at least minimal information about the journal brand.



These patterns show that the main goal of visiting journal websites is to retrieve full-text articles online, although there are different routes to get there. After this goal is achieved, most users terminate their sessions, leaving the journal website.

## II. Format Preference

*PDF is used slightly more than HTML.*

Looking at the ending point of visits and the transition matrix between HTML and PDF, we can conclude that users were slightly more likely to prefer PDF format over HTML format for retrieving full-text articles online.

### A. PDF

PDF format was used slightly more than HTML as the last request of a session. Twenty-six percent (at *Circulation*) to fifty-six percent (at *Journal of Immunology*) of sessions requested PDF after reading HTML versions of full-text articles.

If there is a choice of PDF full text without going through HTML after reading an abstract, users tend to prefer requesting the PDF format directly.

But the evidence of users' format preference can be seen more clearly from sessions referred by PubMed. Nearly all sessions from PubMed in this study are directed to HTML versions of full text, but the majority of users request PDF after reading articles in HTML.

In summary, users seem to like PDF versions, probably because they can leave the web with copies of the articles, either in print or archived in the computer.

### B. HTML

So why do users bother to request HTML for full text retrieval if they can get the PDF version directly? HTML is more friendly for reading articles on the screen, and it seems to be used as an intermediate point—to scan/read articles quickly before printing or archiving them. HTML also provides hyperlinks to cited articles, a feature our user surveys have found users do use and find very useful.

Despite these advantages of HTML, HTML is used not as much as PDF, most likely because HTML is not as print friendly—and the final goal of most journal web visits is printing.

## III. Online Feature Usage

As we have seen, the most common purpose of journal web visits is to print out the article. Thus, requests for online features were a relatively small percentage of usage compared to full-text requests. However, if we look at the absolute numbers (frequencies) of requests for online features, they are significant numbers. We have many online features in the data but here we selected a few features to analyze. Available features vary by journal, so we selected core features available for most of the participating journals. Features we analyzed were hyperlinks to cited articles, access to articles in press, and access to field-specific resources/databases.

### **A. Cross reference to non-HighWire journals/within HighWire journals**

This is a feature providing hyperlinks to cited articles listed and it was the most popular online feature among many features, across all the journals. Links to cited articles are available only through e-journals, and this is the online feature most used and found most useful according our previous survey studies. 1 to 2 % (50 to 2,000 clicks) of total clicks per journal requested hyperlinks to cited articles.

### **B. Hyperlinks to articles in press**

Not all journals we collected web log data for have this specific feature. Hyperlinks to articles in press provide access to manuscripts which are already peer-reviewed and accepted and are in press. Users can read the articles before they appear in the printed edition. This serves the need (particularly acute for life scientists and medical practitioners) to obtain articles/information as soon as possible. Users can read articles without waiting for a new issue to become available on the web or be delivered by post. Among journals that had this feature on February 13, 2000<sup>5</sup>, 1 to 2% (60 to 3,600 clicks) of total clicks requested links to articles in press.

### **C. Field-relevant resources/databases**

One preliminary new finding from the data mining was that users may value “field specific resources/databases” as much as they value retrieving full-text articles online. This feature is available only at the *Journal of Nutrition* website, which provides access to a “nutrient information” database. Users requested this feature as frequently as they requested full-text articles in HTML/PDF. According to usage statistics from this journal, 22% of all clicks requested this feature, while 24% requested full text either in HTML or PDF. And many sessions (18% of total sessions) ended with “nutrient information.” This finding indicates that users value field-relevant information more than other common online features and this type of feature is likely to attract more traffic to journal websites.

## **IV. Search patterns**

Most searches are followed by another search—users start searches and they refine their searches more than once. Sessions with searches contained more clicks than sessions without searches, and spent more time on each request. However, users seemed to use searches to locate articles instead of browsing contents. The most popular component for search commands was search for keyword at “title and abstract”. The next-most popular were by “authors’ names”, and “key word anywhere.” And small but significant numbers of searches were done by “volume and issue numbers,” too. Search patterns were not significantly different between users of medical journals and users of life sciences journals.

## **4. CONCLUSIONS**

In the time e-journals have been available to scholars, journal developers and publishers have been wondering how users behave on the Web. So these suppliers collected data on usage, hoping it would help them understand users’ behaviors better. However, they soon realized that

---

<sup>5</sup> *Journal of Biological Chemistry, Journal of Applied Physiology, The Plant Cell, and Radiology.*

just having the data was not enough to understand users' behaviors, since web log data are so complex to analyze and the scale of the data overwhelming. The data are limited in their ability to describe users experience. One particularly obvious example is the last request of each session: is it the last request because it resulted in success, that is, because it gave the user what was sought? because the user just gave up or ran out of time? because the item last viewed was definitively useless? because the connection failed? We can make some inferences in the aggregate, of course, and our study has tried to find easy ways to analyze web log data to help publishers understand users better. We were not able, however, to discover tools or techniques that could make subsequent analyses convenient or automatic, again, because the data are necessarily massive as well as inherently limited in available data elements. Despite these limits, it is our hope that the following conclusions will be valuable for publishers, librarians, and others in understanding the use of scientific e-journals circa February 2002.

*The final goal of most web visits is a PDF version of an article.*

Most sessions end with a PDF request, and it is well known that PDF is print-friendly. Thus, it is likely that most users visit journal websites with the primary goal of printing out articles of interest. Even users who retrieve HTML versions of articles usually also retrieve the PDF version.

*HTML usage could be increased by making it more print-friendly.*

Our findings suggest that usage of HTML can be boosted by designing HTML to be more print-friendly. For example, currently it is hard to tell how long an HTML article is because there are no page breaks or page numbers. If the ultimate goal of most sessions is printing, users need to be able to see the length of the article immediately in order to make the decision whether to print. If the perks of using HTML—easier reading on the screen and access to hyperlinks—were combined with a print-friendly interface, HTML usage might increase significantly, especially among the many scientific and medical users whose goal is to grab an article to go, rather than to read it in detail on the screen.

*Multi-journal search websites (such as PubMed) create major traffic for journal websites.*

There are two major starting points of visits to e-journals—journal homepages and PubMed—and 30%-60% of sessions were redirected from PubMed. Multi-journal search websites such as PubMed thus play an important role in bringing traffic to journal websites.

*Institutional subscriptions are important to journal traffic.*

Findings show that a significant numbers of sessions (10 to 25%) were referred by academic institutions' websites. This indicates that life scientists used institutional subscriptions to access journal articles. Sessions redirected by PubMed relied on institutional online access to journals especially heavily.

*Hyperlinks to field-specific information/databases may attract users as much as full-text articles.*

As we see from the example of *Journal of Nutrition*, links to field-specific resources, such as Nutrient Information, attract users to journal websites. "Nutrient information" has brought much traffic to the journal website. Web visitors were looking for field-specific information and hyperlinks to "Nutrient Information" appear to be very useful, with 20 to 25% of all clicks requesting this feature. Visitors to *Journal of Nutrition* requested this page as much as they requested full text. Although preliminary (since it's based on only one journal), this finding suggests that journals who represent specific fields could provide this type of feature to attract more visitors to the Web.

Other findings of eJUS<sup>6</sup> suggest that online features in general (including field-specific ones) can be a draw for new personal subscriptions--"I wanted to take advantage of some features the journal had available online" being the third-most popular reason cited for subscribing to a new journal.

*Search patterns don't appear to differ much between users in medical practice and users in the life sciences.*

Search patterns were not significantly different between users of medical journals and users of life sciences journals, despite differences in user characteristics (e.g., the users of medical journals were more likely to be from pharmaceutical companies than the users of biology journals).

Compared to biology journals, medical journals were more likely to have more traffic through journal homepages than through PubMed; however, this can not be generalized with much certainty in this small sample of journals. Some biology journals, such as *Journal of Nutrition*, had more traffic through their homepage than through PubMed.

Our survey studies have shown a number of interesting differences in the scholarly practices of medical practitioners and life scientists. Further research is needed to better understand differences and similarities in usage patterns, including search patterns, between medical practitioners and life scientists.

## **5. IMPLICATIONS**

Despite skepticism from journal developers and suppliers, web log data show that users are fairly comfortable with online journals. On average, users requested 7 to 8 pages on the Web and usually left with full-text articles retrieved online. Given that (according to our survey studies) a main purpose of e-journal use is still to retrieve full-text articles, the fact that 50 to 60% of sessions end with full-text (PDF/HTML) download reflects users' success in getting what they intended on the Web. The actual rate of sessions that obtain full text is much higher (see table A.4), of course, as many sessions continue with other functions after obtaining one or more article texts. E-journals appear to be reaching a fairly mature stage of user adaptation to the new technology.

---

<sup>6</sup> See eJUS<sup>2</sup> second survey report pp. 21 [http://ejust.stanford.edu/findings2/report\\_survey2.pdf](http://ejust.stanford.edu/findings2/report_survey2.pdf)

People visit e-journals through different routes, but multi-journal search websites, such as PubMed, appear to be critical for online journal use. Furthermore, to make PubMed function correctly, institutional subscriptions are very important for journal traffic, since access to journal articles is restricted to subscribers only. This suggests that a strong partnership between publishers and libraries is critical to the future success of e-publishing.

Over the past decade, journal publishers have developed many value-added online features to increase online journal usage. While users are still mainly interested in retrieving full-text articles online, some online features seem to be more successful than others in attracting users to journal websites. Links to field-specific resources/databases on the Web appear to be particularly successful.

This suggests that journals which represent specific fields need to consider providing field-specific resources to users. These features should help journals distinguish themselves from other journal brands. In the scientific publishing world, branding may be the most important factor sustaining revenues from individual journal subscribers. (Scholars typically hold at least one society membership and 1 to 2 subscriptions to journals which represent their own research field, as discussed in the results from our follow-up survey.) Developing hyperlinks to field-specific information/databases on the Web should help journals to distinguish themselves from other journals and stay in business in a highly competitive market. In general, it is likely that any additional services that make useful content readily available through an e-journal web site will increase traffic through the site, which will strengthen the brand identity of that journal as well as increase readership of its articles. To the extent that a journal site is more effective at delivering information for the user (whether or not the content is equivalent to the journal's articles as such), it should help retain or even to grow subscriptions. That is, the e-journal site competes as a discovery and delivery vehicle, as well as competing in the quality of its articles and other editorial content.

## 6. REFERENCES

Cadez, I; Heckerman, D; Meek, C; Smyth P; and White S, 2000., "Visualization of Navigation Patterns on a Web Site Using Model-Based Clustering," Technical Report MSR-TR-00-18, Microsoft Research, Redmond, WA .

Huberman, B. A.; Pirolli, P. L. T.; PitKow, J. E.; Lukose, R. M. 1998. "Strong Regularity in World Wide Web Surfing," Science, 280(3) pp 95-97.

Hand, D.; Mannila, H; Smyth, P. 2001. Principles of Data Mining, MIT Press, Cambridge, Massachusetts

## 7. Appendix

**Table A.1: Transition Matrix given “home page” request: P( • / home page)**

	P(content browse/home page)		P(searches/ homepage)	P(home/ homepage)
	Current	Content		
<b>Radiology</b>	.20	.25	.34	.15
<b>Journal of Nutrition</b>	.19	.17	.32	.15
<b>JPET</b>	.16	.31	.29	.13
<b>J of Immunology</b>	.33	.22	.29	.7
<b>Journal of Clinical Microbiology</b>	.19	.30	.39	.7
<b>Journal of Biological Chemistry</b>		.31	.32	.10
<b>Journal of Applied Physiology</b>	.22		.65	.11
<b>Endocrinology</b>	.21	.34	.32	.10
<b>Clinical Chemistry</b>	.18		.56	.16
<b>Circulation</b>	.25	.25	.27	.12
<b>Chest</b>	.21	.28	.32	.13
<b>Blood</b>	.23	.27	.30	.15
<b>BJP</b>	.14	.35	.29	.17
<b>Plant Cell</b>	.27		.34	.26

Note: Rows of table do not sum up to 1.0 because minor categories included in analysis were omitted in the table. Categories considered are “abstract”, “content-by-date.shtml”, current, shtml, full text in HTML, Home page, Links to cited articles, links to articles in press, PubMed, full text in PDF, and search pages. If transition probability is less than .01 we left out data blank.

**Table A.2. : Transition Matrix given HTML version of full text P( PDF/ HTML)**

	P(PDF / HTML)
<b>British Journal of Psychiatry</b>	.28
<b>Blood</b>	.52
<b>Chest</b>	.41
<b>Circulation</b>	.26
<b>Clinical Chemistry</b>	.34
<b>Endocrinology</b>	.37
<b>Journal of Applied Physiology</b>	.47
<b>Journal of Biological Chemistry</b>	.51
<b>Journal of Clinical Microbiology</b>	.36
<b>Journal of Immunology</b>	.56
<b>Journal of Pharmacology E&amp;T</b>	.53
<b>Journal of Nutrition</b>	.38
<b>Plant Cell</b>	.46
<b>Radiology</b>	.43

Note: Rows of table do not sum up to 1.0 because minor categories included in analysis were omitted in the table. Categories considered are “abstract”, “content-by-date.shtml”, current, shtml, full text in HTML, Home page, Links to cited articles, links to articles in press, PubMed, full text in PDF, and search pages. If transition probability is less than .01 we left out data blank.

**Table A.3. : Host Name**

	Unidentified IP addresses (%)	U.S. Academics (%)	Foreign Countries (%)
<b>British Journal of Psychiatry</b>	20	7	32
<b>Blood</b>	30	13	38
<b>Chest</b>	30	9	44
<b>Circulation</b>	28	12	32
<b>Clinical Chemistry</b>	31	9	38
<b>Endocrinology</b>	27	15	45
<b>Journal of Biological Chemistry</b>	30	23	36
<b>Journal of Clinical Microbiology</b>	31	8	39
<b>Journal of Applied Physiology</b>	24	21	36
<b>Journal of Immunology</b>	27	25	41
<b>Journal of Pharmacology E &amp; T</b>	31	17	34
<b>Journal of Nutrition</b>	26	13	22
<b>Plant Cell</b>	27	15	32
<b>Radiology</b>	32	13	35

Note: Rows do not sum up to 100% because some host codes were not listed in the table. Those omitted host codes are web addresses combined with \*.com, \*.org, \*.gov, or \*.net such as google.com, aol.com, nih.gov, ochsner.ochsner.org, fcrdc-nas-116.NCIFCRF.GOV, and nc-stgermain.noos.net.

**Table A.4. : Sessions requested full text (HTML/PDF) at least once**

<b>Journal Name</b>	<b>HTML (%)</b>	<b>PDF (%)</b>
<b>British Journal of Psychiatry</b>	39	15
<b>Blood</b>	60	37
<b>Chest</b>	43	39
<b>Circulation</b>	68	30
<b>Clinical Chemistry</b>	46	34
<b>Endocrinology</b>	60	38
<b>Journal of Applied Physiology</b>	50	35
<b>Journal of Biological Chemistry</b>	51	65
<b>Journal of Clinical Microbiology</b>	47	47
<b>Journal of Immunology</b>	57	56
<b>Journal of Nutrition</b>	37	18
<b>Journal of Pharmacology &amp; ET</b>	50	45
<b>Plant Cell</b>	38	24
<b>Radiology</b>	50	19