University of Tennessee, Knoxville
Trace: Tennessee Research and Creative Exchange

School of Information Sciences Publications and Other Works

School of Information Sciences

January 2008

Viewing and Reading Behaviour in a Virtual Environment: The Full-Text Download and What Can Be Read Into It

David Nicholas

Paul Huntington

Hamid R. Jamali

Ian Rowlands

Tom Dobrowolski

See next page for additional authors

Follow this and additional works at: http://trace.tennessee.edu/utk_infosciepubs

Part of the Library and Information Science Commons

Recommended Citation

Nicholas, David; Huntington, Paul; Jamali, Hamid R.; Rowlands, Ian; Dobrowolski, Tom; and Tenopir, Carol, "Viewing and Reading Behaviour in a Virtual Environment: The Full-Text Download and What Can Be Read Into It" (2008). *School of Information Sciences Publications and Other Works.*
http://trace.tennessee.edu/utk_infosciepubs/6

This Article is brought to you for free and open access by the School of Information Sciences at Trace: Tennessee Research and Creative Exchange. It has been accepted for inclusion in School of Information Sciences Publications and Other Works by an authorized administrator of Trace: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

Authors
David Nicholas, Paul Huntington, Hamid R. Jamali, Ian Rowlands, Tom Dobrowolski, and Carol Tenopir

This article is available at Trace: Tennessee Research and Creative Exchange: http://trace.tennessee.edu/utk_infosciepubs/6

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0001-253X.htm

Viewing and reading behaviour in a virtual environment

The full-text download and what can be read into it

David Nicholas, Paul Huntington, Hamid R. Jamali, Ian Rowlands and Tom Dobrowolski
Centre for Information Behaviour and the Evaluation of Research (CIBER), School of Library, Archive and Information Studies, University College London, London, UK, and

Carol Tenopir
School of Information Sciences, University of Tennessee, Knoxville, Tennessee, USA

Received 8 January 2008
Revised 18 January 2008
Accepted 26 January 2008

Abstract

Purpose – This article aims to focus on usage data in respect to full-text downloads of journal articles, which is considered an important usage (satisfaction) metric by librarians and publishers. The purpose is to evaluate the evidence regarding full-text viewing by pooling together data on the full-text viewing of tens of thousands of users studied as part of a number of investigations of e-journal databases conducted during the Virtual Scholar research programme.

Design/methodology/approach – The paper reviews the web logs of a number of electronic journal libraries including OhioLINK and ScienceDirect using Deep Log Analysis, which is a more sophisticated form of transactional log analysis. The frequency, characteristics and diversity of full-text viewing are examined. The article also features an investigation into the time spent online viewing full-text articles in order to get a clearer understanding of the significance of full-text viewing, especially in regard to reading.

Findings – The main findings are that there is a great deal of variety amongst scholars in their full-text viewing habits and that a large proportion of views are very cursory in nature, although there is survey evidence to suggest that reading goes on offline.

Originality/value – This is the first time that full-text viewing evidence is studied on such a large scale.

Keywords Transactional analysis, Electronic journals, Information retrieval, Reading

Paper type Research paper

1. Introduction

An increasing number of publishers, librarians and academics are constructing strategies, policies and budgets on the back of the copious volumes of usage data issuing forth from the logs of digital libraries and publisher platforms. Usage data are being variously used to demonstrate customer satisfaction, utility, value for money, and, increasingly, the positive academic outcomes of generous digital library access.

Logs, of course, provide plenty of evidence of use and, to a lesser extent, satisfaction, and the UCL Centre for Information Behaviour and the Evaluation of Research (CIBER)

Aslib Proceedings: New Information Perspectives
Vol. 60 No. 3, 2008
pp. 185-198
© Emerald Group Publishing Limited
0001-253X
DOI 10.1108/00012530810879079

has been a pioneer in extracting a range of robust use (and user) metrics from these data and has encouraged the adoption of a wide range of metrics to measure usage and satisfaction. However, the information community has tended to concentrate on one metric, and that is the full-text view or "download", regarded to be the usage gold standard. The full-text download is generally perceived to be a user satisfaction indicator and, more controversially, a proxy for "reading", so providing the much sought after evidence that a positive academic outcome has taken place as a result of the provision of access to digital information resources.

We believe that this paper evaluates, for the first time, the evidence regarding full-text viewing behaviour by pooling together data on the full-text viewing of tens of thousands of users studied as part of a number of investigations of e-journal databases conducted as part of CIBER's Virtual Scholar research programme.

2. Aims and objectives

Given the importance attributed to full-text viewing behaviour (it is regarded as a metric of satisfaction) and the absence of studies investigating it, there is a clear need to evaluate this form of information seeking behaviour. This paper provides a comprehensive evaluation of full-text viewing behaviour as evidenced in the transactional logs. In this connection it will examine:

- the frequency with which full-text views are made;
- the characteristics of full-text viewing;
- the diversity in full-text viewing, especially in regard to different scholarly groups; and
- the significance of full-text viewing, especially in regard to its role as a proxy for reading.

A thorough investigation of the available literature shows that despite the fact that the metric is widely employed and so much store is set by it, surprisingly, nobody has subjected full-text views to the substantial and robust investigation that has been undertaken in this paper.

The data upon which the paper is largely based come from CIBER's Virtual Scholar programme, an investigation into the information seeking behaviour of academics. The programme has chronicled the behaviour of several hundred thousand scholars, in connection with a variety of digital journal libraries[1], including Blackwell Synergy (Nicholas *et al.*, 2005), ScienceDirect (Nicholas *et al.*, 2008), OhioLINK (Nicholas *et al.*, 2006) and the OUP journal *Nucleic Acids Research* (Nicholas *et al.*, 2007b). For the purpose of this paper we shall largely, but not exclusively, concentrate on the log evidence obtained from the two most recent investigations, that of:

(1) OhioLINK, (2005-2007) an investigation of the use made of the journals database by member institutions, which has created an evidence base of more than 4,000,000 views to 6,000 + journals, an investigation which formed part of the MaxData research project; and

(2) ScienceDirect, which examined usage of nearly 1,500 journals by nearly 800 authors over the 18-month period January 2004 to June 2005.

In regard to the other supporting studies, the Synergy study covered the usage of 671 journals during 17 September 2003. Over a quarter of a million page views were made during this period. The *Nucleic Acids Research* study concerned usage in regard to just this journal: 2,500,000 page views were analysed for the period January 2003-June 2005.

The OhioLINK investigation also featured a questionnaire study undertaken by MaxData colleagues led by Carol Tenopir[2]. One of the purposes of the survey was to obtain an explanation for the viewing habits portrayed in the logs. and this data is referred to selectively in this paper.

3. Literature review

As noted in the Introduction, the importance of usage data monitoring has increased notably during the last few years (Gedye and Park, 2004), and this has been reflected in an increase in usage studies. Some have questioned putting so much importance on usage data. Thus Peters (2002) stated that e-resource usage statistics may not move us any closer to a clear understanding of the real use of information than gate counts and circulation statistics once did. He pointed out that we continue to study indicators of use, not use itself, and that we need to be careful about any inferences made from an analysis of usage data about the needs, interests, and preferences of users. Town (2004) raised the question "why is there now such a focus on usage data, particularly when little or no effort was made to assess use of the equivalent print serials collections in the past?". The answer, of course, is that it is much easier to collect the data.

Usage data has been seen to play a role in budget justification and making decisions about subscription models. The "big deal" in particular has caused much soul-searching, and inevitably usage data has played a part in the debate (Nabe, 2001; Friend, 2003; Ball, 2004). Generally, most studies have shown that "big deals" lead to a considerable increase in usage of journals. Gargiullo (2003) analysed full-text download statistics of an Italian consortium to evaluate the efficiency of their "big deal" subscription model. As usage increased during the study period (2002-2003), she concluded that the "big deal approach definitively seems to be a right choice to make" (p. 298). The British Joint Information Systems Committee (JISC) also funded a COUNTER usage data analysis in 2004 in order to obtain accurate and up-to-date data on the national use of e-journal "big deals" negotiated by the JISC within the NESLi2 initiative. The findings showed a 42 per cent rise in use over 18 months. The study concluded that in terms of cost and usage, libraries were getting good value for money from the e-access fee charged by the publisher (Bevan *et al.*, 2005). On the back of usage data, Obst (2003) concluded that it cost less to use electronic articles than print ones.

The underlying assumption made in most of the above papers is that logs represent real and significant use. Town (2004) questioned this assumption:

> . . . usage data appears to be taken to indicate use, and to be used for evaluation in a way that assumes all usage constitutes useful use. This seems a questionable assumption, given that what is being recorded is access rather than use. This may be positively dangerous if used to inform selection decisions or impute the comparative value of resources. Such decisions require understanding of the context of use and a more definitive statement of value from the user. Counting was said to be "no substitute for listening".

For Davis (2004), the real value of the article download was that it provided robust evidence of the size of the user community. He also warned (Davis, 2007) of the dangers of employing download data:

1. Downloads are not public (like citations) and are therefore open to self-interested abuse, by authors, publishers, or librarians; 2. Downloads are influenced by a publisher's interface; 3. Because of article copies being located in many different places (from the publisher's site, an aggregator's database, a subject, institutional, governmental, or personal archive), aggregating these statistics becomes problematic.

Tenopir and colleagues have pioneered the study of scholars' reading behaviour since the 1970s. In all of Tenopir and King's surveys they defined a "reading" as "going beyond the table of contents, title and abstract to the body of the article". They also stated that the total annual amount of time spent reading had not changed much over the years, although estimates of the reading time spent per article had increased from about 45 minutes per article to 52 minutes. They considered the increase in the average length of articles (from 7.4 pages to 11.7 pages) as a possible reason for this. The amount of reading and time spent reading each article may, of course, vary from field to field (Tenopir and King, 2000). In a recent survey of paediatricians (Tenopir *et al.*, 2007), Tenopir and her colleagues found that they read quickly – on average only 22 minutes per article. The main reason driving their reading was current awareness, and for this they mainly relied on their print subscription rather than electronic versions of journals.

4. Methods

This paper puts a particular methodology – transactional log analysis – under the microscope in respect to the information it provides about a particular information seeking activity – the viewing of full-text articles. A refined and more sophisticated form of transactional log analysis, i.e. deep log analysis (DLA), is used for this purpose. It is called "deep" because it works with the raw server logs and therefore can provide more detailed and complex analyses than can be provided by proprietary or COUNTER-compliant logs. Furthermore, it also enables use data to be furnished in a user context, such as, for instance, in regard to students, as this paper demonstrates. Log data are processed by the Statistical Package for the Social Sciences (SPSS) and subjected to deep log forms of analysis. More details of DLA can be found in Nicholas *et al.* (2007).

There are particular problems in calculating full-text views as counts can be considerably inflated by the fact that if a user comes in from a gateway, like PubMed, they are typically taken to an HTML link and then users often decide to go on to view the article as a PDF. This results in double counting. COUNTER have attempted to overcome this by introducing a rule which says "All double clicks on an http-link within 10 seconds of each other will be counted as only one request. Where a PDF-link is involved, this filter is set at 30 seconds, due to the longer time it takes to render a PDF" (see www.projectcounter.org/). Of course, this has to be a rough and ready measure, because what if the time interval is 11 seconds? The alternative, which was adopted by the authors of this paper in a study of *Nucleic Acids Research*, was to only count unique item downloads (see www.oxfordjournals.org/news/oa_report.pdf).

This paper also uniquely employs time online analyses in order to better understand the character and function of the full-text view. Each page delivered by the server to the requesting client is date-stamped and page view time is calculated as the difference between the time stamps. However, because nobody logs off on the web, the last view in a session cannot be estimated. It should be noted that at least some of the pages that

the user viewed will have not been recorded by the server, and this is because the user would have made use of cached pages during their search. Cached pages are previously viewed pages temporarily stored by the client machine, and for speed subsequent views were from this local store.

All of the Virtual Scholar studies referred to in this paper used exactly the same methodology, DLA, which means it is possible to build a coherent picture of full-text downloading from a number of separate studies.

5. Results

What all the CIBER studies show is that there is a tremendous "activity" associated with digital journal libraries – large numbers of page views, sessions conducted and visits made – and that this activity is increasing. Not only are more people being drawn into the scholarly net, but existing users are searching much more freely and flexibly thanks to laptops, wireless and broadband connections. Significantly, it is also clear that a good deal of this activity arises from the scholar making choices and grappling with the digital information cornucopia. Thus a major CIBER finding has been the discovery of a widespread, pronounced and endemic form of digital information seeking behaviour which is best described as "bouncing" (Nicholas *et al.*, 2007). Bouncing is a form of behaviour whereby a high proportion of users view only a few web pages from the vast numbers available on a site and a substantial proportion (usually the same ones) generally do not return to the same website very often, if at all. It is argued that this form of behaviour is the result of search engine searching, a shortage of time and huge digital choice – factors which combine to create a horizontal rather than vertical form of information seeking (i.e. bouncing).

Hence the interest in full-text viewing, because it has largely been assumed to be a rather more substantial and robust form of use. What follows is an examination of Virtual Scholar log data in regard to:

- the key characteristics of full-text viewing and its diversity; and
- the meaning and significance of the data.

5.1 Full text viewing as a proportion of all viewing

The popularity of full-text viewing can be best gauged by turning to the findings of the OhioLINK investigation of use at four universities over a 15-month period, January 2005 to April 2006 (Nicholas *et al.*, 2007a). In this study, page views were classified into five groups, views to:

(1) menus, which included views to the alphabetical and subject menus of database (journal) content;

(2) lists, which included journal and issue lists;

(3) the search facility;

(4) abstracts; and

(5) full-text articles.

Lists recorded the most views (719,674); they were followed by articles (580,164), the search option (364,713) and abstracts (258,772), with menu items accounting for 176,018 views. The general result therefore indicates that users undertook a wide range

of actions online and viewing full-text articles was just one of them and not the most frequent one at that. Browsing and navigating towards text was clearly a significant form of behaviour in the case of large digital libraries like OhioLINK. Also, with more than a quarter of a million abstracts viewed, this has to be a powerful testament of the enduring popularity of abstracts – perfect, perhaps, for making choices in a crowded digital information environment, and this is especially significant in the case of OhioLINK where users did not have to view an abstract before they could view the full-text, which is typically the case in other digital journal libraries.

Whether a full-text article was viewed or not depended on the status of the user. Thus the Ohio study showed that, when given a choice of viewing an article in abstract or full-text form, students were markedly more likely to opt only to view a full-text article in a session than faculty staff[3]. Possibly, university staff appreciated the brevity and relevance-checking characteristics of abstracts. Thus, 64 per cent of student sessions saw just a full-text article viewed as compared to 50 per cent in the case of the faculty.

5.2 Number of article views in a session

The Ohio study showed that an average of two unique full-text articles was viewed in a session, the actual figure being 2.23. It is probably more helpful to group views by the number undertaken in an online session, and in the case of the OhioLINK half of all users (49 per cent) viewed just one article in a session, over a third (36 per cent) viewed two to four different articles, 10 per cent viewed between five and ten different articles in a session and 2 per cent viewed 21 or more articles.

The Synergy study showed that there was quite a difference between types of users. Take the case of users defined by geographical location: on average, 17 per cent of sessions undertaken by Japanese users saw views to two or more articles in a session. The equivalent figure for Canadian users was just 7 per cent. This might be explained by cultural factors. The ScienceDirect study's contribution was to show that the average number of articles viewed in a session also varied quite considerably between subject fields. Thus materials scientists recorded the highest average of 2.7 articles per session, while medical users recorded the lowest average of just 0.8 articles – quite a difference. Furthermore, this was not simply a function of the number of journals available to each group. Thus, for example, while the number of engineering journals available was about 100 titles more than the number of physics journals, the average number of articles viewed in physics was just 0.1 higher for engineering.

5.3 Relationship between the method of navigating towards content and full-text viewing behaviour

OhioLINK data showed that whether or not a full-text article was viewed during a session depended to a certain extent on the navigational route or mode of access the user took to finding content (Nicholas *et al.*, 2006). In understanding the significance of these data it is worth remembering that, in the case of OhioLINK, users could choose to view an article as an abstract or as full text from the same screen. If the abstract was used, this format must have been regarded as being sufficient to meet the user's needs; equal to or better than an article, because they clearly could have viewed an article for (almost) the same amount of effort. Those sessions employing an alphabetic browsing list to navigate towards content were easily the most likely to view only a full-text article – 74 per cent did so as compared to 56 per cent of sessions where the search

engine was employed. The most likely explanation for this perhaps lies in the fact that search engine users had a greater number of pages to view and they resorted to the abstracts to make a quick selection – abstracts are plainly quicker to load and read and it is much easier to make comparisons.

5.4 Meaning and significance of full-text viewing

It has been demonstrated that a large number of full-text views are made but, of course, not all full-text views are equal, some might represent nothing more than a cursory inspection and rejection of an article, some might constitute a rapid recognition of relevance and download to read offline later, others might involve the reading of just a page or two of the article, and yet others might involve the reading of the whole article online. It is impossible to detect this from conventional usage logs but it is possible to get closer to the truth by looking at:

- the duration of full-text viewing (for instance, is it of sufficient duration to suggest online reading?); and
- repeat viewing, that is whether the full-text article was viewed more than once, either in the same format or in a different format (i.e. first viewing an article in HTLM and then requesting a PDF).

5.5 Duration of an online full-text view

As mentioned previously, bouncing points to the adoption of a horizontal rather than vertical form of information seeking, and this is confirmed by the time online analysis. This is because the online viewing times for full-text articles are very short indeed, even allowing for problems in determining this outlined in the Methodology section. Thus, while the OhioLINK study of four universities showed that of all the page viewing opportunities open to them, unsurprisingly, users spent longer viewing an article than anything else, the actual time spent was not that significant. Employing Huber's M-estimator to overcome a skewed distribution, menus took eight seconds to view, lists 17 seconds, the search screen 25 seconds, abstracts 24 seconds, and articles 106 seconds (less than two minutes). Two thirds of article views lasted three minutes or less and an astonishing 40 per cent were completed in a minute or less. What this tells us is that a large proportion of full-text online views were extremely brief and, possibly, cursory.

This finding is confirmed by the ScienceDirect study, where articles also took the longest to view. However, the average median time of about 38 seconds was around a third the time recorded for OhioLINK. The explanation for the considerable difference could be that the users were less likely to be students, and we shall learn in the next section that students were more likely to read online and spent more time doing so than faculty.

The Ohio study showed marked differences between faculty and students in their full-text viewing habits and students were likely to spend more time viewing an article online. Thus in terms of average (median) page time staff viewed articles more rapidly, 77 seconds as compared to 106 seconds for students, quite a considerable difference in online information seeking terms. Also, while 9 per cent of staff recorded a view time of between seven and half minutes and one and half hours (sufficient time, perhaps, to read an article online) this was true of 14 per cent of students.

The supporting questionnaire survey, which asked people in four Ohio universities about the last article they read (not viewed), in what form and how long it took them to read it provides some useful context, and some support, for the log data. It was found that faculty members said they took about quarter of an hour to read an article completely online and that the maximum recorded reading time value was two hours. It was also found that students said they took slightly longer reading online, about 20 minutes (median), with the longest taking about four hours.

The Ohio study also showed that there were big differences between subject fields. Chemists and life scientists recorded the longest article viewing times – 106 and 112 seconds, respectively – while computer scientists (55 seconds) and business economists (60 seconds) recorded views of about half that time. These findings were supported by the aforementioned questionnaire data, which showed that the discipline in which users were most likely to read electronically was science.

A comparison of four Ohio universities showed that there were significant differences between them. Thus the median estimate of article view time was significantly longer at the two research universities (89 seconds for them both) as compared to 76 and 68 seconds for the two teaching universities, which is perhaps what one might expect.

The ScienceDirect study additionally found that scholars spent relatively more time viewing shorter (4-10 pages long) articles online than longer (21 + pages) articles. Neither of the times, 42 and 32 seconds respectively, suggested that anyone was really doing more than rapid scans online, but clearly relatively more time was spent on shorter articles. This would suggest that shorter articles obtain relatively more attention online, and that, maybe, had a better chance of actually being read.

It was also discovered that scholars might be trying to avoid reading online. Thus, as the length of a journal article increased, there was a greater likelihood that it would be viewed as an abstract and less likelihood that it would be viewed in full text (Nicholas *et al.*, 2008). Thus, while 84 per cent of articles less than ten pages long were viewed as either a PDF or HTML, this figure fell to 73 per cent for papers 11-20 pages long and to 58 per cent for papers 21 or more pages long.

The format in which an article is displayed has an impact on viewing time. The study of Synergy showed that, although PDF files took longer to download, HTML files recorded a longer view time by over half as much again – 142 seconds as compared to 94 seconds (Huber's M-estimate) for PDFs. An explanation for this might lie with the fact that PDFs are more difficult to read on screen and, although some users may well be reading the PDF on screen, it seems more likely that PDF viewed articles were being printed or later consulted offline. HTML formatted articles are easier to read and manipulate on screen, and although some users will print them off in an HTML format it seems likely that users – those recording an above average HTML view time – were probably reading the article online. The choice of format may well vary with institution. Thus users in some locations – like a busy library – may not be able to work comfortably "online", and hence opt to print out a PDF formatted file. There may be a cost implication too – a PDF formatted file is likely to be cheaper to print than an HTML file as the PDF formatted file can be usually displayed on fewer pages.

5.6 Identifying online reading behaviour

In the case of the Ohio study the proportion of full-text views that might constitute online readings was estimated. Informed by the previously mentioned findings of the

questionnaire survey, which showed how long staff and students said they took to read an article completely online, the Ohio logs were examined for those articles with an online viewing time falling within these possible reading reference time points. Two time points were considered:

(1) five minutes and two hours; and

(2) a smaller range of seven and a half minutes to one and a half hours.

The smaller time span was selected as it would be expected that fewer read online articles would have either a value lower or higher in the range with most articles being read having a median value (see Table I).

In terms of the views falling between five minutes and two hours the median read value of these articles was just under 13 minutes (767 seconds), which falls below the expected median of between 15 and 20 minutes as estimated by the questionnaire. Taking the slightly shorter period, seven and half minutes to one and half hours, the estimated average (median) reading time was about 17 minutes (1,019 seconds) and falls in the middle of the faculty and student median as estimated by the questionnaire data. However, this reduces the percentage of articles being read online from 17 per cent to 12 per cent – quite a difference. In part this is the result of the setting of the lower boundary. At the point of five minutes this may well include users who are not reading the whole document; perhaps, these users were cherry-picking, reading just one or two sections or have taken time in deciding to print the article out.

It was also decided to set the article timeout signal at the 90 per cent percentile of those reading an article between 7.5 minutes and one and half hours, which was 56 minutes (3,356 seconds). Of course it was not possible from looking at the logs to determine whether an article was read or printed. There will be instances where a user would have, say, printed an article off and then terminated their session. We would not be aware of the termination and the recorded article time would be high. In this case the article will have been misclassified as having been read and the session would have been deemed terminated after a 56-minute lapse. However, this error is thought to be a lesser error compared to setting the timeout signal low. This would have been the case by taking the 90 per cent percentile, a value of about 22 minutes (1,321 seconds), of all article view time. Taking this value would have increased the risk of imposing an inappropriate session end of those who read an article online.

5.7 Repeat viewing

The Ohio study shed light on repeat article viewing. The viewing of the same article more than once could be seen as a sign of interest and relevance. Most (77 per cent) articles were viewed once in a session, 12 per cent of sessions saw one article viewed again and 11 per cent of sessions saw at least one article viewed two or more times or

Table I. Estimates of those reading an article online by time range (in seconds)

	Percentage of articles	Mean	Median	10 per cent percentile	90 per cent percentile
Five minutes to two hours	17.1	1,391.7	767	349	3,548
7.5 minutes to 1.5 hours	12.2	1,490	1,019	515	3,356

two articles viewed more than once. It was also shown that research active universities were less likely to examine an article again in a session; perhaps, their users were more practised and so able to determine what they wanted in the first place.

Perhaps a stronger indicator of interest would be if the same document was viewed in different formats. Thus if someone first views a document in HTML format and then requests it in PDF format – something that is quite obvious from the logs, then this would provide evidence that this was a relevant document.

A repeat viewing analysis was conducted as part of a usage study of *Nucleic Acids Research*. Very high levels of repeat viewing were identified, which was partly explained by the fact that users coming in from the NCBI and PubMed gateways are taken to the HTML version of the article requested, and to view the PDF version the user has to come out of the HTML version and load up the PDF version. This process resulted in the downloading of two items rather than one. To give an idea of the scale of this, Figure 1 gives the distribution for views in a session of an item just as a PDF version or as a HTML version or as both by referrer organisation (grouped).

Thus, for users coming to the journal via either NCBI or PubMed, no one viewed just a PDF version; they only viewed a PDF version after they had first viewed an HTML version. The majority of users (59 per cent) coming in from NCBI viewed both a PDF and HTML full text version, and 41 per cent just viewed an HTML text. The same was true of users coming into the site via PubMed – 41 per cent had viewed both a PDF and HTML version and 49 per cent had just viewed an HTML version. These findings probably say much about the effectiveness of the search facilities on the gateways. By comparison only one in eight search engine users viewed an article again.

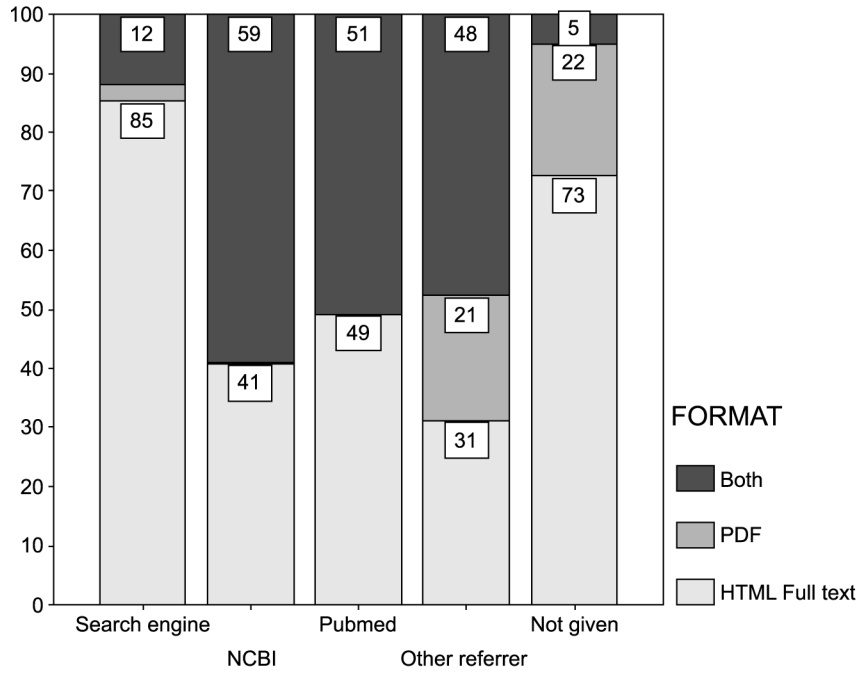

Figure 1.
Nucleic Acids Research: views in a session of an article just as a PDF, as a HTML version or as both by referrer organisation (grouped)

6. Discussion

The data that have been presented demonstrate clearly the need to look beyond the headline full-text download data in log reports, because the data are quite diverse and raise important questions as to what we can read into full-text viewing. So why then are publishers, librarians and many researchers seemingly so content with headline full-text counts? In the case of librarians it is possible that they are simply not aware that there is a problem because their COUNTER-compliant reports, through which most acquire their usage data, do not provide the necessary level of analysis and accuracy. COUNTER was never set up to perform the kind of bespoke *user* analyses presented in this paper.

Another possible reason for the full-text download's enduring popularity is the cost and complexity of moving to a more sophisticated full-text metric as outlined in this paper. In particular, making time calculations in an environment in which individual users are hard to identify, typically do not log off and search for a whole variety of purposes is clearly difficult, although possible.

Determining the extent to which full-text views provide a reliable indicator of interest, value or satisfaction is problematic. Thus, while the evidence indicating that users do not spend much time on viewing a full-text document and that the bigger the document the less time they will spend on it would appear to suggest that, in many cases, they are not. But this would neglect two important points. Firstly, the articles may be read later, either in print or digital form. This was borne out by the questionnaire survey of the faculty and students of four Ohio universities. They were asked what was the final form in which they read their last article. "Reading" was defined as going beyond the table of contents, title, and abstract to the body of the article. In the case of faculty, downloaded and printed on paper was mentioned most frequently (43 per cent), followed by print article in print journal (29 per cent), online on computer screen (14 per cent), photocopy (9 per cent) and then previously downloaded/saved and read, on computer screen (6 per cent). This tells us several important facts:

- about two-thirds of last readings were of articles identified online; and
- a high proportion were printed out for later reading, and only one in six full-text views were actually read while online.

From this it can be estimated that 63 per cent of readings related to online full-text views. Of course, what the Ohio data does not tell us is how many views were made before it was decided to read an article online or download or print it for later on, so these figures have to be regarded as being on the low side.

Secondly, users might get all they needed from viewing just a paragraph or page (they might have been looking for a fact), and/or they were gathering data from a number of sources as they navigated the digital information space.

7. Conclusions

The paper has undertaken a comprehensive examination of full-text viewing for the first time, and while it cannot be claimed to have ended arguments about the academic outcomes of this form of behaviour, it has set out the chief characteristics of this form of behaviour, demonstrated how diverse this form of behaviour is, and put forward a method of distinguishing between types of full-text viewing.

The main log findings follow:

- Virtual scholars undertook a wide range of activities when online, and viewing full-text articles was just one of them and not necessarily the most popular. Navigating towards content in very large digital spaces is the user's chief concern as demonstrated by the number of views to menus, lists and search pages. This is where much of the user's energy is expended, and much of this is the result of what we may loosely call power browsing, moving rapidly through the digital space picking information up on the move. Navigating is not a secondary activity but a fundamental one.
- Some user groups were more likely to view the full text of an article and this was particularly true of students, who would not have had as good access to scholarly communications as members of staff and would have to face print charges if they wanted to read offline.
- Some forms of information-seeking behaviour resulted in higher levels of full-text viewing. This was particularly true of the method of navigating chosen towards content, where people using an alphabetic browsing list proved to be the most likely to record a view just to a full-text article.
- Typically, articles were viewed for, on average, less than two minutes, insufficient time to read them fully, although enough time to power-browse them.
- There were big differences in terms of the time users spent viewing a full-text article: scholars at research universities spent longer viewing an article than their counterparts in teaching universities, suggesting more interest being shown by the former group; students spent longer viewing an article than staff, suggesting a greater predilection to read online; of the disciplines, chemists and life scientists recorded the longest article view times.
- Shorter articles received relatively more time spent on them and, maybe, had a better chance of being read online.
- Scholars might be trying to avoid reading too much online. Short viewing times tell us this, but so does the finding that as the length of a journal article increased there was a greater likelihood that it would be viewed as an abstract and less likelihood that it would be viewed in full text.
- The online information-seeking process dictates that you print or download first and then take the decision about relevance later. Therefore, inevitably, a good number of full-text downloads will never be read – it constitutes an action of "sending to printer".

Further, from the supporting questionnaire data from the OhioLINK study, the following was discovered:

- In the case of both faculty (63 per cent) and students (73 per cent) the vast majority of readings were generated from online searching.
- Students were far more likely to read online than staff, a quarter of their readings were of this type. This was a finding which confirmed the findings of the log studies.

- For students the median reading time (for all articles online) was 20 minutes, somewhat higher than that for staff, which was 15 minutes. This too confirmed log data.
- Triangulating log and questionnaire findings it was found that just 12 per cent of full-text views could be assumed to have resulted in online reading.

Finally, the results of this study, when taken together with CIBER research on "bouncing" (Nicholas *et al.*, 2007), show more than anything else that rich navigating opportunities and the power browsing that goes with it are the prized benefits and real outcomes of web searching. To focus on full-text views and cling on to notions of reading is to miss the point, which is that we have witnessed a paradigm shift in the way that people use information systems. The major activity that pre-digital information users undertook was reading, they had limited and confined information choices and these were often made for them (by librarians) and, as a result, the opportunities and tools for navigation were limited also. Today, with massive digital choice, a high information churn rate and an absence of information filters (indeed, the individual is the filter) the main user activity is inevitably navigation, and we should not be surprised by this, or indeed believe that navigation itself is not fulfilling or informative in itself. The most important feature of today's information environment is links and this has made information seeking a horizontal rather than vertical form of behaviour. It is thus impossible to isolate reading from navigating, people are reading as part of searching, not searching for reading.

Notes

1. For brief background information about each of these research projects, see www.publishing.ucl.ac.uk/research.html
2. See the web site: web.utk.edu/ ~ tenopir/maxdata/project_docu.htm
3. The results should be treated with caution because the definition of staff/faculty was based on the labels given to sub-networks used.

References

Ball, D. (2004), "What's the 'big deal', and why is it a bad deal for universities?", *Interlending & Document Supply*, Vol. 32 No. 2, pp. 117-25.

Bevan, S., Dalton, P. and Conyers, A. (2005), "How usage statistics can inform national negotiations and strategies", *Serials*, Vol. 18 No. 2, pp. 116-23.

Davis, P.M. (2004), "For electronic journals, total downloads can predict number of users: a multiple regression analysis", *Portal: Libraries and the Academy*, Vol. 4 No. 3, pp. 379-92.

Davis, P.M. (2007), "UKSG usage factor research – an update", e-mail discussion to toliblicense-l@lists.yale.edu (accessed 11 March 2007).

Friend, F.J. (2003), "Big deal-good deal? Or is there a better deal?", *Learned Publishing*, Vol. 16 No. 2, pp. 153-5.

Gargiullo, P. (2003), "Electronic journals and users: the CIBER experience in Italy", *Serials*, Vol. 16 No. 3, pp. 293-8.

Gedye, R. and Park, T. (2004), "COUNTER – increasing the value of online usage statistics", *Serials*, Vol. 17 No. 2, pp. 155-8.

Nabe, J. (2001), "E-journal bundling and its impact on academic libraries: some early results", *Issues in Science and Technology Librarianship*, No. 30, available at: www.istl.org/01-spring/article3.html (accessed 20 January 2008), Vol. 20.

Nicholas, D., Huntington, P. and Jamali, H.R. (2007a), "Diversity in the information seeking behaviour of the virtual scholar: institutional comparisons", *Journal of Academic Librarianship*, Vol. 33 No. 6, pp. 629-38.

Nicholas, D., Huntington, P. and Jamali, H.R. (2007b), "Open access in context: a user study", *Journal of Documentation*, Vol. 63 No. 6, pp. 853-78.

Nicholas, D., Huntington, P. and Jamali, H.R. (2008), "User diversity: as demonstrated by deep log analysis", *The Electronic Library*, Vol. 26 No. 1, pp. 21-38.

Nicholas, D., Huntington, P. and Watkinson, A. (2005), "Scholarly journal usage: the results of deep log analysis", *Journal of Documentation*, Vol. 61 No. 2, pp. 248-80.

Nicholas, D., Huntington, P., Jamali, H.R. and Dobrowolski, T. (2007), "Characterising and evaluating information seeking behaviour in a digital environment: spotlight on the 'bouncer'", *Information Processing and Management*, Vol. 43 No. 4, pp. 1085-102.

Nicholas, D., Huntington, P., Jamali, H.R. and Tenopir, C. (2006), "What deep log analysis tells us about the impact of 'big deal'. Case study OhioLink", *Journal of Documentation*, Vol. 62 No. 4, pp. 482-508.

Obst, O. (2003), "Patterns and costs of printed and online journal usage", *Health Information and Libraries Journal*, Vol. 20 No. 1, pp. 22-32.

Peters, T.A. (2002), "What's the use? The value of e-resource usage statistics", *New Library World*, Vol. 103 Nos 1/2, pp. 39-47.

Tenopir, C. and King, D.W. (2000), *Towards Electronic Journals: Realities for Scientists, Librarians and Publishers*, Special Libraries Association, Washington, DC.

Tenopir, C., King, D.W., Clarke, M.T., Na, K. and Zhou, X. (2007), "Journal reading patterns and preferences of pediatricians", *Journal of the Medical Library Association*, Vol. 95 No. 1, pp. 56-63.

Town, S. (2004), "E-measures: a comprehensive waste of time?", *Vine*, Vol. 34 No. 4, p. 191.

Corresponding author
David Nicholas can be contacted at: david.nicholas@ucl.ac.uk

To purchase reprints of this article please e-mail: **reprints@emeraldinsight.com**
Or visit our web site for further details: **www.emeraldinsight.com/reprints**