A WEB LOG ANALYSIS OF THE ONLINE *NSW PUBLIC HEALTH BULLETIN* FOR 2001–2003

Johanna I. Westbrook
Health Informatics Research & Evaluation Unit
University of Sydney

D. Lynne Madden
Editor
NSW Public Health Bulletin

ABSTRACT

The web logs of the online version of the *NSW Public Health Bulletin* were analysed to understand the patterns of use. Twenty-nine months of data, for the period January 2001 to May 2003, were extracted from archived files stored by the NSW Department of Health. HTML and PDF hits were included; other types of hits, for example image hits, were not. Five potentially useful variables were identified: Internet protocol address; date of access; time of access; document accessed; and means of access. There were 384,887 hits during the period, approximately 442 per day. The rate of hits per month increased from 8288 in 2001 to 21,288 in 2003. The PDF version was used more than the HTML version. Examination of HTML hits revealed how different parts of the *Bulletin* were being used. This information provides evidence to inform planning.

The *NSW Public Health Bulletin* was established in May 1990 as part of the newly developed public health infrastructure in New South Wales. Its purpose is to enable the timely communication of information on public health issues and thus to contribute to the development of a well-trained and informed public health workforce.

A printed copy of each issue is distributed to a wide range of public health workers in a variety of settings. It is also provided online via the NSW Health Department's website www.health.nsw.gov.au/public-health/phb/phb.html. A PDF version has been available since the mid 1990s and in September 2001 an HTML version was launched as part of a new *Bulletin* home page. At that time all the issues for 2001 that had been published were made available in HTML. In late 2003 the authors undertook a web log analysis study of the *Bulletin* to better understand who uses the online version and how frequently they use it.

BACKGROUND TO WEB LOG ANALYSIS

Use of a web site is usually measured by web server logs, which automatically record access to a website. These files automatically record user identification information in the form of an Internet protocol (IP) address. IP addresses are registered by organisations. Some IP addresses are useful for providing an indication of the origin of those using the website. However, the level of detail of IP addresses varies considerably. For example, it is possible to identify IP addresses registered to universities and to some specific health care organisations, but a large proportion of IP addresses are registered to private Internet providers and for a proportion of IP addresses no organisation can be identified. In addition to IP addresses, web logs routinely store information about the time and date of access and some information about the documents that were viewed on the website.

Analysis of web logs can provide useful information regarding the identity of users of specific websites and when and how users seek out information from those sites. The value of the analysis of web logs is largely dependent upon the level of detail of information recorded in the logs. In general, web logs provide massive amounts of data but are limited in the amount of detail and precision they provide.[1] As Nicholas et al wrote when commencing the analysis of web logs of the online version of *The Times* newspaper in Britain ' ..nothing can prepare you for the sheer size of the [web log] datasets and their propensity to grow' (p266).[1] Previous studies using web log analysis to investigate the search behaviours of people using online library catalogues and knowledge databases have been undertaken and demonstrate both the strengths and weaknesses of this approach in answering specific research questions.[2-9]

The first step in log analysis is to determine what definition of use will be adopted. The most commonly used measure is 'hits' to a website—a hit being defined as a unit of information, delivered from the server to a browser, that makes up part of a web page access. Thus a hit may be either a text hit or a graphic hit. Web logs are not able to identify individual users unless users are required to enter a unique identifier. Hits provide a comparative and not an absolute measure of utilisation. Their value lies in answering questions such as, is use generally increasing or decreasing, or is some content more popular than other content.

METHODS

Data

Twenty-nine months of web log data for the period January 2001 until May 2003, relating to the *Bulletin*, were extracted from archived files stored by NSW Department of Health. Only HTML and PDF hits were included in the analysis. Image hits, for example, were removed because an image hit is recorded for every picture and diagram included in an article. Thus an article with several images will record multiple hits in the log file (one for the text and several for the pictures associated with the article). Removing these image hits from the web log dataset provides a more accurate representation of the frequency with which specific articles are accessed. Figure 1 shows an extract from the log data file.

FIGURE 1

EXTRACT FROM THE *NSW PUBLIC HEALTH BULLETIN* WEB LOG DATA FILE

158.232.66.185 | 27/May/2003 | 18:23:11 | +1000 | "GET | /public-health/phb/HTML2002/aug02html/worldreport.html | HTTP/1.1" | 200 | 16189 | "-"
66.77.73.77 | 27/May/2003 | 18:32:15 | +1000 | "GET | / public-health/phb/jan01html/Guestedtjan01.html | HTTP/1.0" | 200 | 16391 | "-"
203.12.140.120 | 27/May/2003 | 18:32:29 | +1000 | "GET | /public-health/phb/phbjuly02.pdf | HTTP/1.0" | 200 | 249503 | "www.google.com.au/search?hl=en&lr=&ie=UTF-8&q=accidental+death+inequality+and+the%27+aborigines%27&spell=1" 210.84.35.169 | 27/May/2003 | 18:46:17 | +1000 | "GET | /public-health/phb/phbsubj.html | HTTP/1.1" | 200 | 904720 | "www.health.nsw.gov.au/_living/travel.html"144.138.242.93 | 30/May/2003 | 18:15:11 | +1000 | "GET | /public-health/phb/HTML2002/july02html/renaldisease.html |

The data were cleaned and additional programming undertaken to improve the value of the data for analysis. For example, a specific script was developed to map pathways from IP addresses (represented by numbers in the logs) to their named users, allowing us to identify the sites of specific organisations such as private Internet providers and universities through which users were accessing the *Bulletin*.

Content of the *NSW Public Health Bulletin* web logs

The dataset contained five potentially useful variables in terms of answering questions related to by whom, how, when and how often, the electronic version of the *Bulletin* was used (Table 1). Figure 2 shows the steps required to access content within an issue of the *Bulletin* in both the PDF and HTML versions.

Individuals arrive at the *Bulletin* home page and select either the current or back issues option. They are then given the option of viewing the entire *Bulletin* as one document (the PDF version) or viewing individual articles (the HTML version), which they access by clicking on the table of contents and then selecting a specific article for viewing. With the HTML version, if the user wishes to view another article in the same issue they are required to return to the table of contents and select the article. The web log stores the web address as each selection is made. Thus when the HTML version is selected the web log records a hit for the table of contents for that issue and a hit for each specific article viewed.

When a user selects the PDF version of an issue they are able to scroll through all articles within that issue. The web log will record only one hit, signifying that the PDF file for the issue was accessed.

The total number of HTML hits per issue does not, therefore, reflect the number of viewers of that issue of the *Bulletin*, as each HTML user will on average produce two to three hits. So the number of HTML hits is two to three times higher than the number of people who have viewed the HTML version of that issue. PDF users generate only one web log hit per issue viewed.

Assumptions and analysis

Some assumptions about the data were made during analysis and interpretation of the results. It was assumed that HTML users view one article on each visit to the site. Therefore, once the user has accessed the *Bulletin* home page and selected an issue, on average every HTML user will generate two web log hits in order to view an article, whereas a PDF user will generate only one hit and has access to the entire issue's content. It was decided,

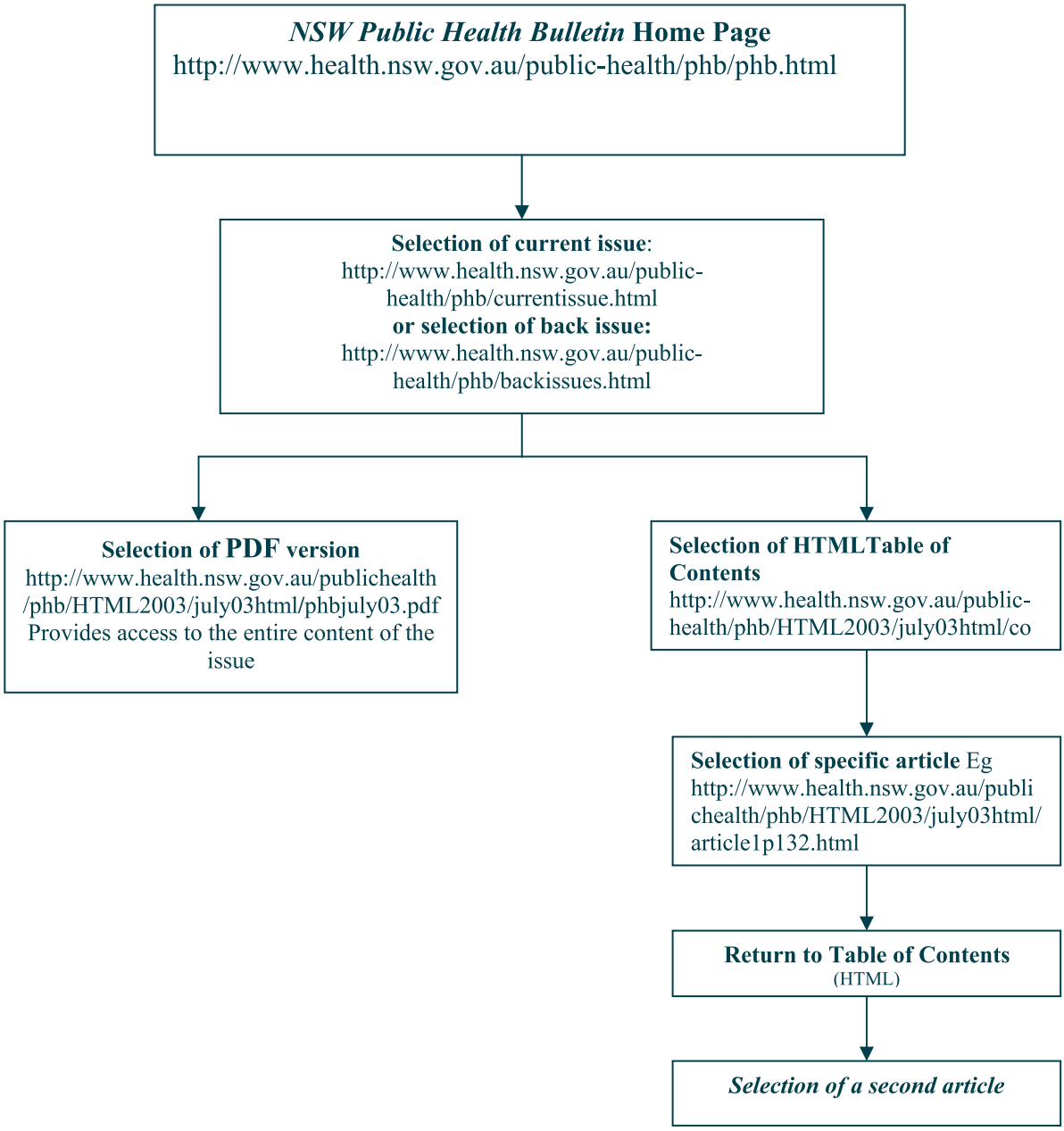
TABLE 1

VARIABLES OF INTEREST IN THE WEB LOGS OF THE *NSW PUBLIC HEALTH BULLETIN*

Variable	Detail
Internet protocol (IP) address	Indicates origin of user. Analysis of this variable is limited as many IP addresses are generic (eg searches undertaken via public internet providers such as BigPond, Primus etc) and a proportion of IP addresses cannot be identified. Categories of users that can be identified include those originating from specific universities, and users from outside Sydney via information in their web address (eg Hunterlink). Users from specific countries overseas may also be identified if their country of origin is specified in their web address eg '.au', '.uk' etc).
Date of access	Day, month, year
Time of access	Hours and minutes
Document accessed	This indicates the web address of the document viewed. It indicates whether the document was HTML or pdf. Documents related to specific issues of the NSW Public Health Bulletin can be identified and the nature of the content (eg Fact Sheet) is sometimes apparent. This variable is dependent upon the way in which each page was named and some inconsistencies in naming over time were apparent.
Avenue through which the searcher reached the NSW Public Health Bulletin	For example, via the NSW Department of Health home page, or a search engine such as Google.

FIGURE 2

STEPS TO ACCESS CONTENT WITHIN AN ISSUE OF THE *NSW PUBLIC HEALTH BULLETIN*



therefore, that reducing the HTML hits by 50 per cent would provide a more accurate indication of the popularity of the HTML version compared to the PDF version.

To gain access to a particular issue of the *Bulletin*, readers must access the *Bulletin* home page and then select current or back issues. Each of these hits is also registered in the log file. In order to assess the use of the HTML and PDF versions, all these 'background' hits were removed.

To investigate the extent to which users viewed specific regular sections within the *Bulletin,* hits to these documents were examined. This analysis was only possible where users had selected the HTML version of the *Bulletin*. The analysis assumed that the same labels were used for these articles in every issue of the *Bulletin*. Searches for hits to the following specific documents were performed: the Communicable Diseases section (Search on label = 'commdis'); and Fact Sheets (Search on label = 'facts').

The total number of hits and rates of hits per month, year and issue were calculated.

Data quality issues

Some inconsistencies in the labeling of the HTML and PDF documents were detected. For example, some of the HTML Fact Sheets were identified in terms of the issue and year, while others were labelled according to the topic of the Fact Sheet.

The Communicable Diseases section was usually labeled 'communicable diseases' but in the Jan/Feb issue for 2003 the section was labeled ' www.health.nsw.gov.au/public-health/phb/HTML2003/janfeb03html/janfebommdiseasesreport.html' and thus hits to this document were not initially detected using the search string above. Wherever possible these inconsistencies were identified and addressed in the analysis.

RESULTS

Web utilisation patterns for the *NSW Public Health Bulletin*

In total there were 384,887 hits to the *Bulletin* during the 29 months reviewed. This averaged 13,272 hits per month, or approximately 442 per day. Rates of hits per month increased from 8,288/month in 2001 to 14,690/month in 2002 and 21,288/month in 2003 (over the five months of data available for 2003). Figure 3 shows that hits to the *Bulletin* website increased considerably over the study period. These data represents when hits occurred but does not reflect whether readers were seeking information from current issues of the *Bulletin*, or from back issues.

Use was greatest at the beginning of the week and lowest on the weekend (Figure 4).

Forty-nine per cent of use occurred between the hours of 9 am and 5 pm and 80 per cent occurred on weekdays. The 10 per cent of use occurring between 1 am and 3 am (Figure 5) may reflect access from people overseas in a different time zone.

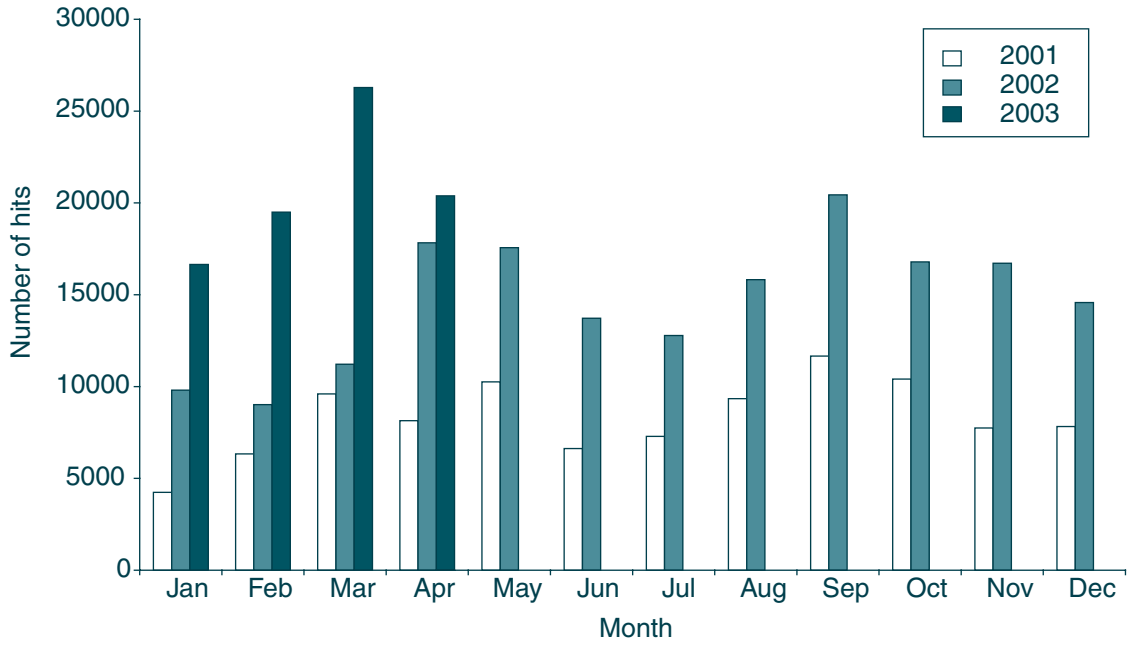
Pattern of use on each of the weekdays was similar (Figure 6), while times of use on Saturday and Sunday varied (Figure 7).

HTML versus PDF use

Figure 8 shows that in each year of the study period, the PDF versions of the *Bulletin* were accessed around two and a half times more frequently than the HTML versions (based on the assumption that each PDF hit on the *Bulletin* is equivalent to two HTML hits, as explained in the Methods Section). The lower percentage of HTML hits in 2001 is most probably explained by the fact that the HTML version of the *Bulletin* was first made available in September of that year. HTML versions of all issues of the *Bulletin* published in 2001 were put on the web that September.

FIGURE 3

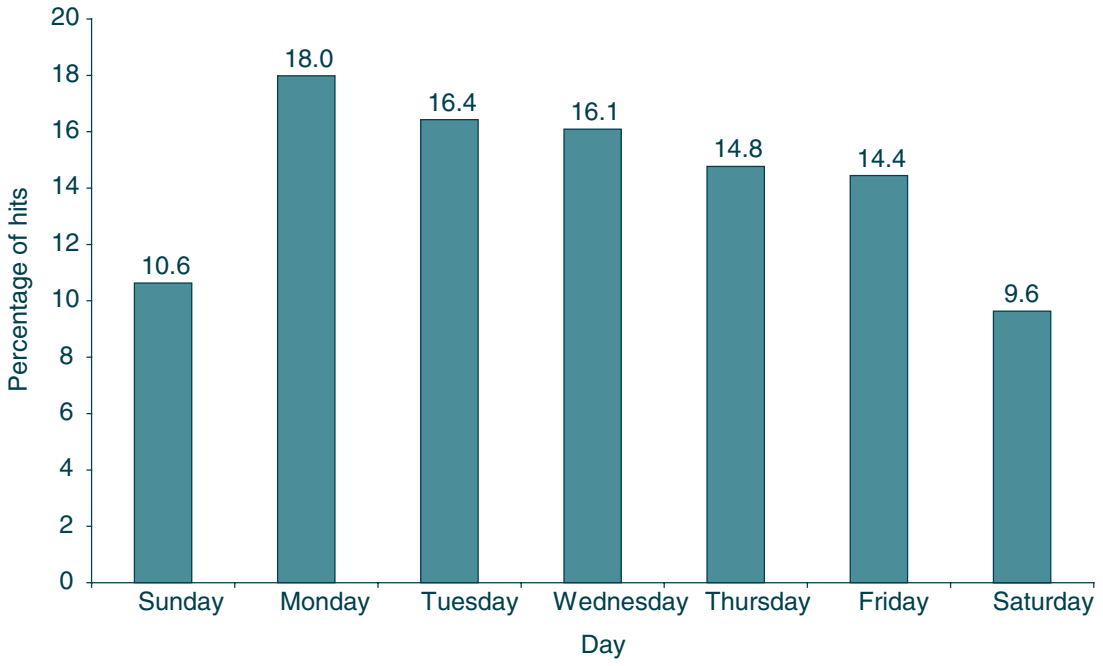
MONTHLY HITS TO THE NSW PUBLIC HEALTH BULLETIN WEBSITE, JANUARY 2001 TO APRIL 2003



Source: NSW Department of Health web log archives

FIGURE 4

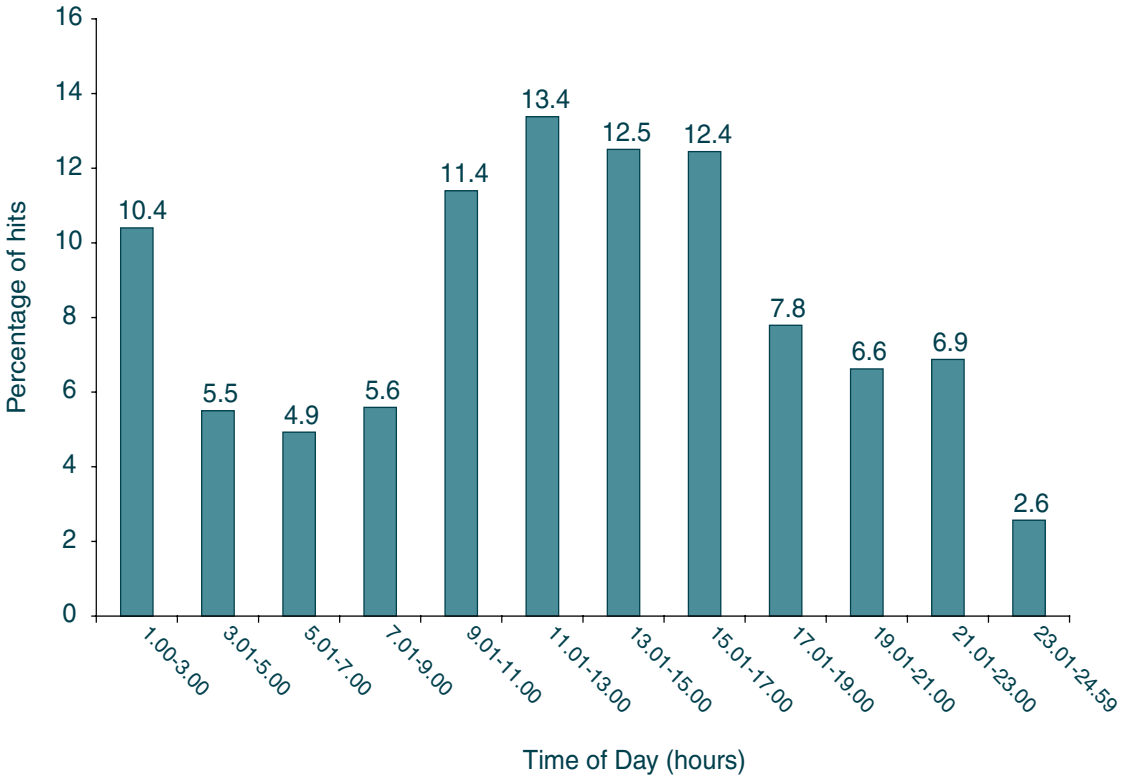
HITS TO THE NSW PUBLIC HEALTH BULLETIN BY DAY OF THE WEEK



Source: NSW Department of Health web log archives

FIGURE 5

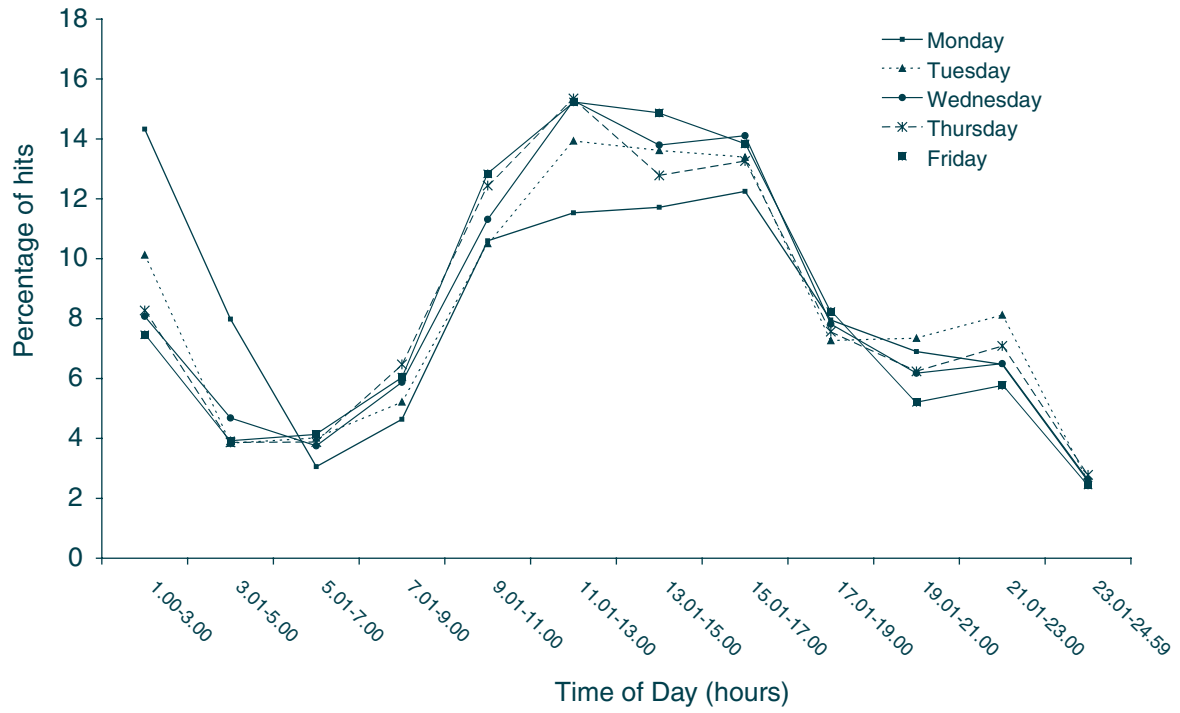
HITS TO THE NSW PUBLIC HEALTH BULLETIN BY TIME OF ACCESS



Source: NSW Department of Health web log archives

FIGURE 6

HITS TO THE NSW PUBLIC HEALTH BULLETIN ON WEEKDAYS, BY TIME OF DAY



Source: NSW Department of Health web log archives

Identification of users

For 28 per cent of hits to the *Bulletin*, no registered organisation could be linked to the associated IP address. In total, 7.7 per cent of hits originated from universities and 6 per cent from the NSW Health Intranet. Twenty-one per cent of hits originated from websites with 'au' in the address, indicating they originated in Australia. However, these do not constitute all hits from Australia, as many Australian web addresses do not have 'au' in them. One per cent of hits originated from the United Kingdom though, again, this is likely to be an under representation as not all UK web addresses have 'uk' in them. For 6.2 per cent of hits the user found the *Bulletin* site via a Google search.

Content accessed

In order to examine specific *Bulletin* content accessed by readers, a subsample of the web logs, consisting of all HTML hits, was extracted. When these data were examined by year of publication, around 8–9 per cent of HTML hits to issues in 2001 and 2002 could be attributed to readers viewing the Fact Sheets. There were not sufficient data for 2003 to estimate this percentage. Figure 9 shows the number of hits to individual Fact Sheets during the 29-month study period. Hits to the Communicable Diseases Report represented 1.3 per cent of total HTML hits in 2001, 2.5 per cent in 2002 and 2.6 per cent in 2003. The Fact Sheets were around three times as popular as the Communicable Diseases section.

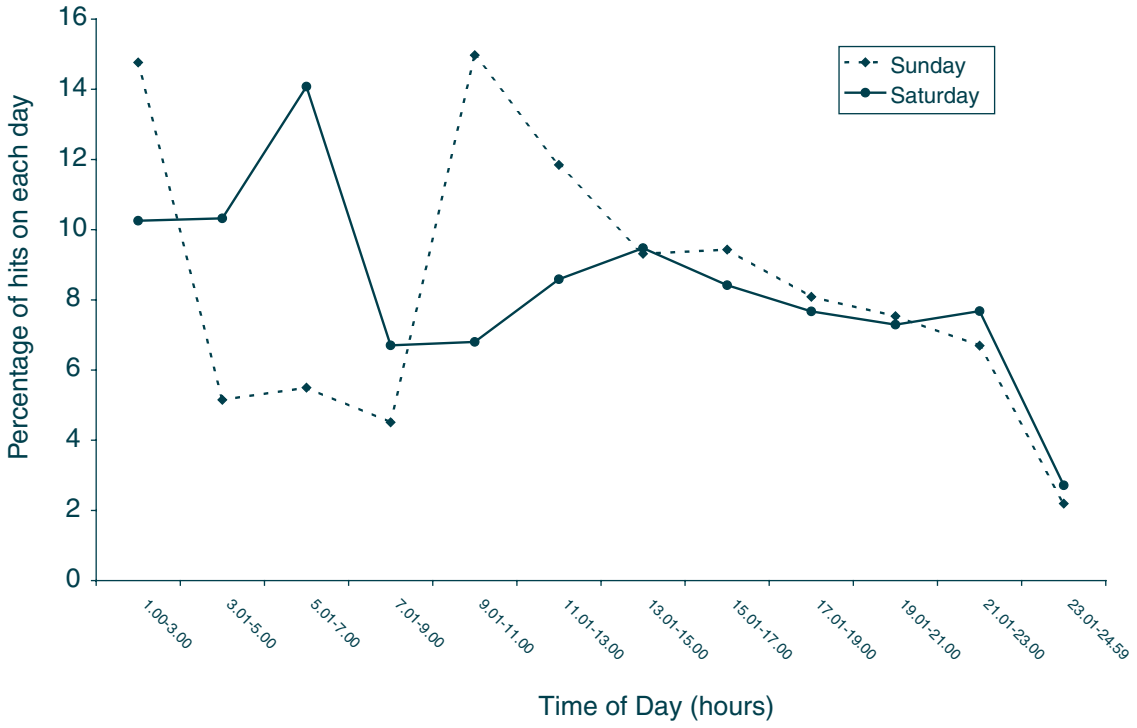
DISCUSSION

There was a considerable increase in hits to the *Bulletin* over the study period. By mid-2003 the volume of hits had more than doubled from those in 2001. Factors that may have contributed to the increase in access to the *Bulletin* include the development of the new home page and production of an HTML version in 2001, and the inclusion of the *Bulletin* in Medline and Index Medicus, which occurred in mid-2002. It was not possible to identify patterns of use of the *Bulletin* for individuals and thereby determine the size of the pool of people who access the *Bulletin*, or the frequency with which they seek information. For example, users may constitute a core group of individuals, each of whom accesses the *Bulletin* on multiple occasions; alternatively, users may consist of a large group who access information only once or twice. The growth in hits to the *Bulletin* could therefore be due to an increase in the pool of users or to an increase in the frequency with which each user seeks information.

Patterns of use in terms of days and times of the week suggest that use is likely to be related to users' work activities, with around 50 per cent of hits occurring between 9 am and 5 pm and 80 per cent occurring on weekdays.

FIGURE 7

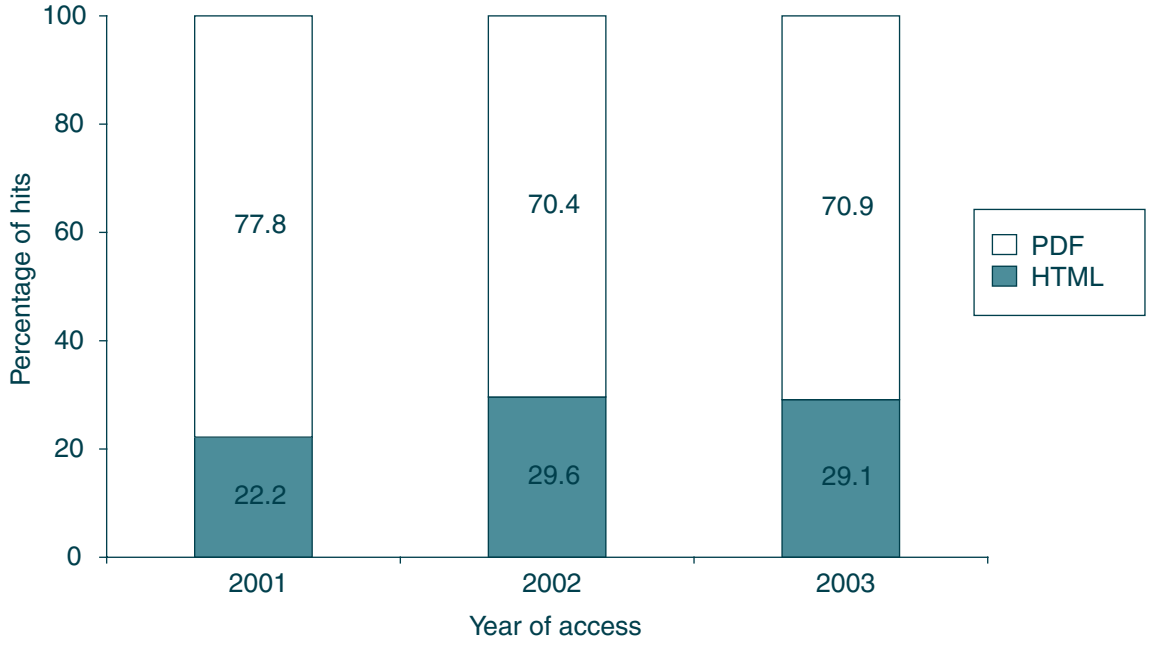
HITS TO THE NSW PUBLIC HEALTH BULLETIN ON THE WEEKEND, BY TIME OF DAY



Source: NSW Department of Health web log archives

FIGURE 8

PROPORTIONS OF HTML AND PDF HITS TO THE NSW PUBLIC HEALTH BULLETIN WEBSITE BY YEAR OF ACCESS*

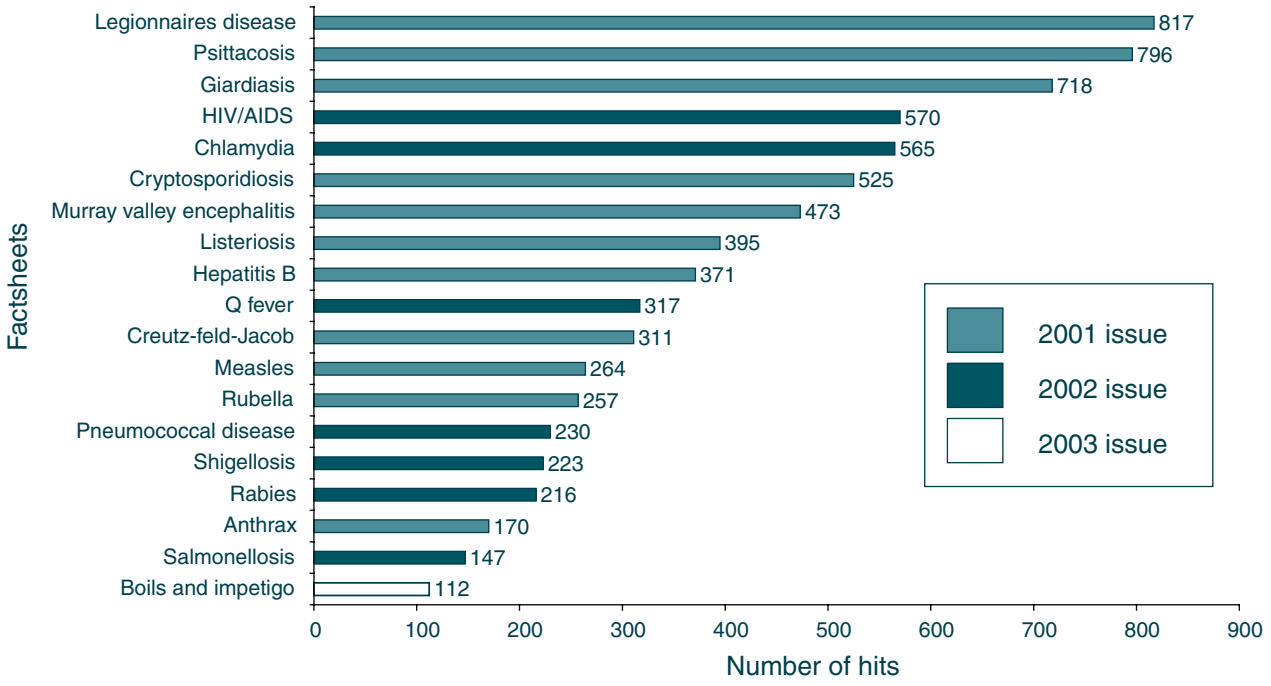


Source: NSW Department of Health web log archives

*The proportions have been adjusted on the basis that accessing the HTML version involves at least two hits and PDFs only one

FIGURE 9

FREQUENCY OF HTML HITS TO SPECIFIC *NSW PUBLIC HEALTH BULLETIN* FACT SHEETS 2001-2003



Eighteen per cent of use occurred between 11 pm and 5 am, which may reflect use by users located overseas and therefore in different time zones.

Conclusions regarding the popularity of specific content within the *Bulletin* were primarily based upon analyses of the HTML users as these individuals select specific content that is then recorded in the web logs. Thus the total number of hits is modest and does not include those users who would have viewed this content via PDF. It would seem reasonable to assume that the type of articles selected for viewing by HTML users is representative of the content read by those who select to view the PDF version or indeed those who read the paper version. If the HTML version is discontinued we will not be able to use the existing web log system to track the use of specific content within the *Bulletin*.

The web logs provided some information regarding the extent to which users accessed regular sections of the *Bulletin*. The results showed that hits to the Fact Sheets make up around 8 per cent of hits to issues published in each year. Those on the subjects of Legionnaires Disease and psittacosis were the most popular. This can partly be attributed to these Fact Sheets being published in early 2001, thereby having a greater time opportunity to attract hits. However, this pattern did not follow for all Fact Sheets. For example, Fact Sheets relating to HIV/AIDS and to Chlamydia were among the top four most popular, yet were published in mid and late 2002.

The Communicable Diseases section of the *Bulletin* did not appear as popular as the Fact Sheets. However, due to the inconsistencies found in the labeling of the Communicable Diseases section for individual issues it is possible that hits to this content were under-estimated. Development of standards regarding the labeling of specific HTML and PDF documents would facilitate the analyses of future *Bulletin* web logs.

Based on the assumptions specified in the methods section, the PDF version of the *Bulletin* is around two and a half times more popular than the HTML version. However, it was not possible to determine whether the PDF and HTML users constitute different populations. For example, individuals may choose to initially use the PDF version, providing access to all content in an issue, and then go to the HTML version at a later date when they wish to quickly locate and print a copy of a specific article or Fact Sheet. Alternatively, individuals may have strong preferences for either HTML or PDF and rarely use the alternative document version. Questions regarding individuals' preferences and use of the *Bulletin* could more satisfactorily be answered using focus groups or a survey.

CONCLUSION

This study allowed the online use of the *Bulletin* to be described in detail for the first time. This information is difficult to obtain by other means, for example by readership surveys that usually have low response rates,

particularly for free publications. The information gained has been used to inform the development of the *Bulletin* website and content.

REFERENCES

1. Nicholas D, Huntington P, Lievesley N, Withey R. *Cracking the code: Web log analysis*. Online and CD-Rom review 1999; 23:263-269.
2. Chisnell C, Dunn K, Sittig D. Determining educational needs for the biomedical library customer: An analysis of end-user searching in MEDLINE, In *World Congress on Medical Informatics* (8th), Vancouver, Canada, International Medical Informatics Association, 1995.
3. Wyly B. *From access points to materials: A transaction log analysis of access point value for online catalogue users. Library Resources and Technical Services* 1996; 40:211-236.
4. Wallace P. How do patrons search the online catalogue when no one's looking? Transaction log analysis for bibliographic system design. *RQ* 1993; 33:239-252.
5. Hunter R. Successes and failures of patrons searching for online catalogue at a large academic library: a transactional log analysis. *RQ* 1991; 30:395-402.
6. Kaske N. Research Methodologies and transaction log analysis: Issues, questions and a proposed model. *Library Hi tech* 1993; 42:79-86.
7. Borgman C, Hirsh S, Hiller J. Rethinking online monitoring methods for information retrieval systems: from search product to search process. *Journal of the American Society for Information Science* 1996; 47:568-583.
8. D'Alessandro M, D'Alessandro D, Galvin J, Erkonen W. Evaluating overall usage of a digital health sciences library. *Bulletin of the Medical Library Association* 1998; 86:602-609.
9. Westbrook J, Gosling AS, Coiera E. Do clinicians use online evidence to support patient care? A study of 55,000 clinicians. *Journal of American Medical Informatics Association* 2004; 11:113-120.