Conference papers

The Primary Care Electronic Library (PCEL) five years on: open source evaluation of usage

Judas Robinson BA Postgraduate student

Simon de Lusignan BSc MSc MD MRCGP Senior Lecturer

Primary Care Informatics, St George's – University of London, London, UK

Patty Kostkova MSc PhD

Research Fellow, City eHealth Research Centre, Institute of Health Sciences, City University, London, UK

ABSTRACT

Background The Primary Care Electronic Library (PCEL) is a collection of indexed and abstracted internet resources. PCEL contains a directory of quality-assured internet material with associated search facilities. PCEL has been indexed, using metadata and established taxonomies. Site development requires an understanding of usage; this paper reports the use of open source tools to evaluate usage. This evaluation was conducted during a six-month period of development of PCEL.

Objective To use open source to evaluate changes in usage of an electronic library.

Method We defined data we needed for analysis; this included: page requests, visits, unique visitors, page requests per visit, geographical location of users, NHS users, chronological information about users and resources used.

Results During the evaluation period, page requests increased from 3500 to 10 000; visits from

1250 to 2300; and unique visitors from 750 to 1500. Up to 83% of users come from the UK, 15% were NHS users. The page requests of NHS users are slowly increasing but not as fast as requests by other users in the UK. PCEL is primarily used Monday to Friday, 9 a.m. to 5 p.m. Monday is the busiest day with use lessening through the week. NHS users had a different list of top ten resources accessed than non-NHS users, with only four resources appearing in both.

Conclusions Open source tools provide useful data which can be used to evaluate online resources. Improving the functionality of PCEL has been associated with increased use.

Keywords: digital, evaluation studies, family, information dissemination, internet, libraries, open source, physicians, primary health care

Introduction

The Primary Care Electronic Library (PCEL) is a directory of abstracted information linking to over 1300 web resources relevant to primary care. The choice of material for inclusion in the PCEL directory is made by members of the PCEL team, who have both clinical and academic interest in primary care. Users may also submit sites for inclusion in the directory,

although these will be vetted by the PCEL team. Each resource is assigned Medical Subject Heading (MeSH) terms and added to the PCEL custom directory.² All of the information and metadata collected regarding a given website is presented to users in the form of an index card. The topics and contents of these index cards are very varied, ranging from the British National

Formulary to the Royal College of Midwives to World Health Organization Disease Outbreak News. There is a bias towards the inclusion of UK-orientated primary care information, although this by no means excludes international resources. One specialist area where PCEL indexes more resources than would be expected is medical informatics: this represents the special interest of the Primary Care Informatics group at St George's, with almost 400 listings hierarchically below the MeSH 'Information Science'.³

Data in the PCEL directory may be searched or browsed in a number of different ways. The directory search allows free-text searching of the contents of the directory, and the advanced search provides more options for advanced users. There are two indexing systems used in the custom directory: one designed by the PCEL team; the other is MeSH. Either system can be browsed to help users discover desired resources. The index cards can also be browsed numerically and alphabetically. Each index card also links to Harvard and Vancouver references of the resource. Rich Site Summary (RSS) feeds, an extensible mark-up language (XML) schema for sharing data between computers, has been developed to describe the ten most recently added resources, and each index card alone. 4

The conceptual thinking behind the PCEL has its roots in the Doctor's Desk project and more recently the Primary Care National electronic Library for Health (NeLH-PC), part of the larger NeLH programme.^{5–7}

The Doctor's Desk project started in 1997 and had to overcome the technical difficulties associated with connecting to practices; it made electronic resources available in practices that had hitherto only been available in academic institutions. The NeLH-PC provided a larger range of resources and these were categorised and indexed using an in-house system. At the end of the NeLH-PC pilot it was decided not to have a separate primary care virtual branch library. Permission was given for NeLH-PC to be left online as PCEL. From November 2004 it has been modernised and is set to continue as a research project. The principal improvements made were:

- automated system of link checking introduced to achieve better connectivity
- online indexing facility added, enabling users to submit resources
- Harvard and Vancouver format references for each index card that can be readily downloaded by users citing that resource
- reindexed using MeSH terms to enable hierarchical browsing of categorised resources.²

Evaluation of digital libraries requires effective analytical tools. Open source distributors provide software to perform this evaluation. The advantage of open source is that it is cheap, at the same time as often resulting in performance improvement. When software is open source, its source code, documentation

St George's University of London	Primary Care Electronic Library (PCEL) Community Health Sciences St. George's, University of London Cranmer Terrace, London, SW17 ORE Email: <u>pcel@sgul.ac. uk</u>
Home Site Map Help My PCEL	Feedback Team PCI
PCEL Directory PCEL has indexed 1378 web sites relevant to Primary Care. in a number of ways:	. Add to these when you <u>submit a site</u> . Directory content can be browsed
Directory search Advanced search Search	By category Browsed via MeSH directory
	Abnormalities
	Numerically or Alphabetically
Biomedical Informatics at SGUL	ABCDEEGHIJKLM
	NOPQRSIUVWXYZ
PCEL accepts <u>submissions</u> to its directory, and if you choose	to fill out personal details you will be accredited for these submissions.
	is site. Individuals may build a custom directory structure containing links
Search PubMed of Go Cle	ear Google Search

Figure 1 The Primary Care Electronic Library (PCEL)

and other content are publicly accessible by acquiring an open-source licence: this does not mean that the software is free. What is freely accessible is the source code: the human-readable instructions for the program. Open source software can be analysed and if necessary altered; it is completely transparent. This is important in log file analysis software (the application used to interpret the web server's log or activity file) as it is an area of computing which has been characterised by misunderstandings, complications and even obfuscations. Part of the difficulty with log analysis arises because common terminology is used whilst differing technical specifications are applied. Definitions of the terms used in this study are shown in Box 1.

We used open source tools to evaluate any changes in usage of PCEL while it was being upgraded between November 2004 and April 2005. This evaluation reports the change in usage over this period.

Method

Data required for the evaluation

We defined the data needed for the evaluation. We decided that despite the lack of a reliable common definition we would include hits. However, we consider page requests, visits, unique visitors and page requests per visit much more reliable measures of

usage (see Box 1). In addition we felt it was important to determine the geographical location of our users, in particular whether they were located in the United Kingdom (UK). We also wished to identify National Health Service (NHS) users as nearly all UK primary care is provided by the NHS. We were interested in what times of day and days of the week the library was used on, and about resources used. Finally, we wanted to be able to identify people who had gone from one of the internet search engines to one of PCEL's index cards, as a link from a busy search engine could distort the apparent level of use of PCEL.

Time period of the analysis

We carried out the analysis of the log files for nine months. The analysis started in August 2004, three months before the programme of site improvements started. We felt that three months of 'pre-intervention' data would give an indication of whether there were trends in the use of the site independent of the site improvements.

Technical aspects

We used Apache, an open source internet server, because it is reported to be the most popular and hosts nearly 70% of current internet sites. ^{11,12} We used log files to record the level of use of the Apache server, and consequently of the PCEL. These records are

Box 1 Common definitions for log analysis

- Hits: each request to the web server is counted as one hit. Hits include a wide variety of file types, including graphics. When a single page is called, a large number of hits may be recorded as multiple requests for graphics files may result. Reports of the number of hits are the least useful of common web statistics as it is next to impossible to correlate them with meaningful events. Although hits are recorded by the package that we used, it is probably best to disregard them. It would not be expected that the number of hits recorded by differing log analysis software analysing the same log files would be the same.
- Page requests (or pages): this records the number of pages of HTML requested by users. This does not include graphic files or the like, and is a more meaningful statistic than hits, representing the text of HTML files transferred to the browser. Pages of HTML are typically identified by the suffix of the file, although differing software may record suffixes differently. We used this measure to gauge the accuracy of open source software packages.
- **Bandwidth**: for each request the log file records the number of bytes of data served with that file. The bandwidth represents the quantity of data served, or the sum of the bytes recorded for each request.
- **Unique visitors**: broadly speaking, visitors can be identified by their Internet Protocol (IP) number which is recorded in the internet log file. Counting the number of unique IP numbers that occur indicates the number of unique visitors to the site. This is regarded as the most accurate measure of traffic. ⁹
- Visit (or session or user session): a visit is a chronologically defined set of requests by a visitor. The cut-off point for a visit is one hour. Thus if a given visitor visits the site twice within a two-hour interval, it will be counted as two visits. In some respects visits are a better indicator of total site activity than unique visitors since they indicate frequency of use.

generated by all internet servers and need to be interpreted by log file analysis software. The log file analysis software that we initially used was called Analog. However, Analog does not report unique visitors, nor does it identify visitors who have arrived via search engines. We therefore moved to another open source product called AWStats to provide this functionality. This log file analysis software cleans log files by excluding requests from common search engines and filtering out requests from the Internet Protocol (IP) numbers associated with development computers. To test whether AWStats' analysis was compatible with Analog we decided to compare the page request numbers generated from the same log files.

Analysis

We carried out the following analyses:

- 1 Comparison of results from Analog log file analysis software with AWStats to see if there were any differences in requests.
- 2 Page requests, visits and unique visitors were measured and graphed against time. We were particularly interested in seeing whether there was an increase in unique visitors as these represent people coming back to the site.
- 3 Page requests per visit: this would provide an indication of how many pages are visited per visit.
- 4 UK and NHS site usage: we were interested to know if our prime target group, UK NHS users, were using the site more. A rise in page requests in either group would imply increased usage.
- 5 Chronology: we also needed to know who uses the site at what times and on which days. This could

- also provide face validity as to whether the proportion of UK users is likely to be correct, as there are few other English-speaking countries in this time zone.
- 6 Use of resources: we wanted to know what the library's most used pages and resources were, and if different resources were popular among NHS users compared with other users.

Results

To illustrate log file parsing, 58% of the total requests for the month of April 2005 were excluded as they were identified as coming from search engines. The two log file analysis programs produced similar but not identical results. Table 1 shows the comparison of page requests for AWStats and Analog. Over the nine months observed, the difference between the programs varied between 0.1% and 12%. Both AWStats and Analog reported a doubling of page requests over the period of the evaluation.

Figure 2 shows the number of unique visitors accessing the site per month, the number of visits, and the number of page requests per month.

Increases in these three measures of popularity can be seen over the period from November 2004 to April 2005, the six-month period over which improvements were made. The number of page requests increased from just over 3500 per month to over 10 000 per month. The number of unique visitors rose from 744 in August 2004 to 1496 in April of 2005. During the same period the number of visits rose from 1240 to 2337. Dips in activity can be seen for Christmas and Easter.

Table 1 Comparison of page requests reported by AWStats and Analog

Month	Awstats	Analog	Percentage difference
August 2004	11 986	11 742	2.1
September 2004	12 815	12 543	2.2
October 2004	12 857	12 244	5.0
November 2004	17 331	17 313	0.1
December 2004	21 451	19 398	10.6
January 2005	19 595	17 509	11.9
February 2005	20 981	19 136	9.6
March 2005	23 502	21 288	10.4
April 2005	25 153	22 716	10.7

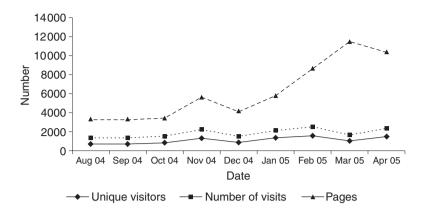


Figure 2 Page requests, visits and unique visitors

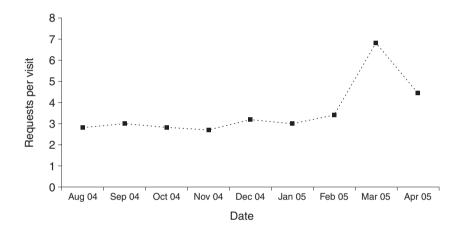


Figure 3 Page requests per visit

The differences in scale in the increase of unique visitors, visits and page requests are illustrated by Figure 3, which shows the number of page requests per visit over the nine-month period.

Page requests per visit can be seen to be steadily increasing (with anomalous data for March, caused by an unrecognised search engine, 'become.com', visiting the site numerous times). The increase in requests per visit is encouraging as it indicates people requesting more pages from the site and a trend away from casual visitors.

Users are more likely to come from the UK than any other geographical location. This is shown in Figure 4, using data for April 2005.

Data for countries is derived from the IP numbers, which are recorded in the access logs. A reverse look-up yields a domain name (such as gtw-13.nhs.uk) that can be analysed for a country suffix. Using this method produces an underestimation of the figure for the percentage of visitors from the UK. Using methods that are capable of tracing UK-based .com domains yields locational percentages of UK visitors as high as 83%. The true figure is probably somewhere in between.

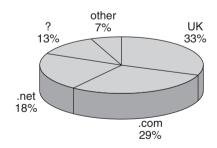


Figure 4 Geographical registration of hosts – April 2005

Figure 5 shows the increase in page requests over the nine-month period under investigation; this has not been accompanied by a significantly increased percentage of international visitors. This is encouraging, as it suggests that PCEL is maintaining a professional UK base.

NHS users can be identified by gateways used to access the internet. Figure 6 shows the total number of page requests, and also the percentage of total page requests, originating from NHS users.

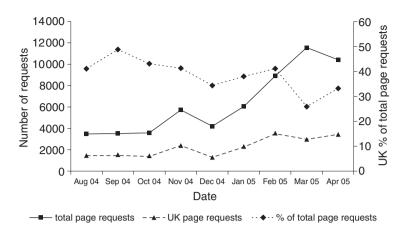


Figure 5 Page requests and percentage of total page requests of UK origin

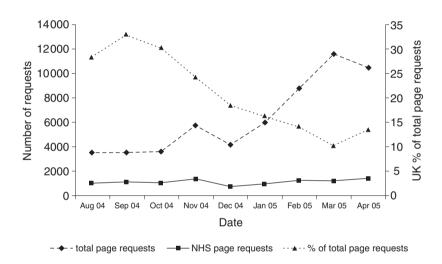


Figure 6 Page requests and percentage of total page requests originating from NHS users

As PCEL has become busier over the period under study, the percentage of NHS users has declined. This does not indicate, however, that NHS use has declined: if anything the total number of page requests from NHS users has increased, but not in proportion with the overall increase in page requests.

On average, in April 2005, the library received 441 page requests per weekday compared with 150 page requests per day at the weekend. This threefold difference is shown in Figure 7.

Between 9 a.m. and 5 p.m. the library received 60% more requests per hour than outside working hours (see Figure 8).

The popularity of the site during the hours of nine to five on working days indicates strongly that the site is reaching its target audience of professionals involved in primary care in the UK.

Tables 2 and 3 compare the most commonly accessed pages of PCEL by all users with those pages accessed by NHS users during April 2005. Of the top ten pages of all users and NHS users for April 2005, seven were present in both lists: these are marked in bold. Browsing can be seen to be vastly more popular than searching for both groups of users, with 1% using the directory search but with over 10% of page requests for browsing the MeSH directory, the custom directory or the alphabetical list of sites. The most frequently requested pages are the home page and the index card which contains details of the resources presented by PCEL. The requested frequency of these two pages is inverted for all users and NHS users, with all users requesting the index card more frequently than the home page and the opposite being true for NHS users. This difference is probably to do with

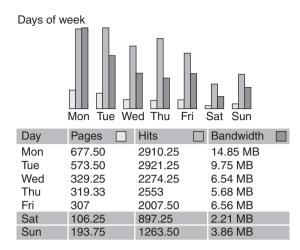


Figure 7 Requests per day of the week for April 2005

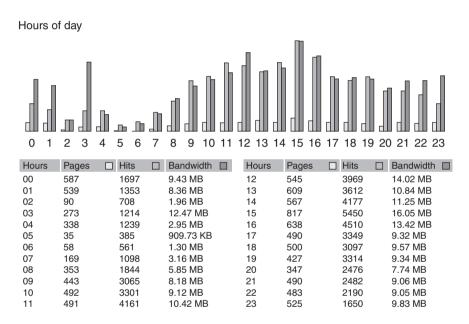


Figure 8 Requests per hour of the day for April 2005

differing use of search engines to find the site. Search engines such as Google have a tendency to point towards the index cards, and if we assume that all users are using search engines such as these more frequently than NHS users, this would account for the difference.

As well as being able to identify the pages accessed by users, we were able to identify the resources accessed by these pages. Tables 4 and 5 show the most commonly accessed resources of the PCEL index card for all users and for NHS users during April 2005. Of the top ten resources for all users and NHS users for April 2005, four were present in both lists: these are marked in bold. Of the remaining six resources there is a bias towards primary care in the selections of NHS users, three of the six being specifically related to primary care (National Vision User Group [NVUG], Torex User Group [TUG] and Guide For Registered Medical

Practitioners [IB204]), compared with one of the six for all users (NHS GP 'Golden Hello' Scheme). Not surprisingly, this suggests a more focused professional interest on the part of NHS users.

Discussion

This study demonstrated that during improvements in the service provision of a digital library, usage increased. The baseline usage of the PCEL and the increase observed over the six months of this evaluation demonstrate the demand for a UK-based primary care internet resource. Almost half a gigabyte of data was transferred by the PCEL in April 2005. Although

Table 2 Most commonly accessed pages	s – all users – April 2005
--------------------------------------	----------------------------

Pages	Viewed	Percentage	
Index card	2702	26.2	
Home page (external link)	1227	11.9	
Browse MeSH	1126	10.9	
Alphabetical list of sites	420	4.1	
Browse custom directory	358	3.5	
Harvard reference	196	1.9	
Vancouver reference	192	1.9	
Home page (internal link)	156	1.5	
Directory search	130	1.3	
RSS feed	66	0.6	
Others	3733	36.2	

Table 3 Most commonly accessed pages – NHS users – April 2005

Pages	Viewed	Percentage	
Home page (external link)	298	21.5	
Index card	193	13.9	
Alphabetical list of sites	66	4.8	
Browse MeSH	44	3.2	
Browse custom directory	38	2.7	
Directory search	11	0.8	
Home page (internal link)	10	0.7	
Site map	9	0.6	
Guideline search	5	0.4	
Newly added sites	5	0.4	
Others	709	51.1	

small in comparison with professionally funded sites, this represents a significant volume of data. Open source tools provide reliable methods for monitoring digital libraries. Using these tools, it can be seen that the usage of PCEL has increased without decreasing NHS or UK usage. The library seems to have broadened its user base without altering its professional make-up. This is demonstrated by the specific preference of NHS users for resources relevant to primary care.

Digital libraries have a role in helping clinicians manage the ever-expanding volume of information available to them. 16

There are limitations to the method. Identification of the country of origin of users was arrived at by a

reverse domain name server (DNS) look-up on the IP numbers recorded in the log files which can yield a country-specific suffix as part of the domain name. Such a method will always underestimate the percentage of UK users, as some UK service providers (such as British Telecom) will resolve to US .com domains. A familiarity with the more advanced features of IP location and the required PERL modules that are part of the AWStats package would probably solve this problem. The second limitation of the methodology applies to restricting the AWStats program to specific hosts, such as .edu or .ac.uk. This restriction, which would provide detailed information on types of visitor, is only available if DNS look-up is already

Table 4 Most commonly	v accessed	resources – al	l users – April 2004	5
Table 4 Wiest Commission	y accessed	i coodi cco di		•

Resource	Viewed
Wheeless' Textbook of Orthopaedics	66
ECG Library	57
Superficial Thrombosis	40
Making the Best Use of a Department of Clinical Radiology: Guidelines for Doctors	37
ACP Journal Club	28
Department of Health – Essence of Care	26
Systematized Nomenclature of Medicine (SNOMED)	23
Journal of Neonatal Nursing	21
Information Technology in Nursing	19
NHS GP 'Golden Hello' Scheme	19

Table 5 Most commonly accessed resources – NHS users – April 2005

Resource	Viewed
Wheeless' Textbook of Orthopaedics	12
Journal of Neonatal Nursing	9
Making the Best Use of a Department of Clinical Radiology: Guidelines for Doctors	8
ECG Library	6
National Vision User Group (NVUG)	6
TUG (Torex User Group)	5
International Classification of Diseases (ICD)	4
Medical Defence Union	4
Guide for Registered Medical Practitioners (IB204)	4
DermIS Dermatology Online Atlas	3

done in the log file, which is not the case in our set-up. Thirdly, the capture of NHS use is not all-encompassing. NHS professionals, for whatever reason, may not use NHSnet gateways to access the internet and to access PCEL. A further limitation of the research is that all search engine traffic may not have been excluded by the default filters on the log analysis software. Thus, although general conclusions about usage may be drawn, more specific inferences are beyond the scope of this study.

We chose to use Apache's internet server; an alternative would have been Microsoft's internet information server. These two together have 90% of the market share. There is little difference between the two. In fact, despite '... the array of differences between the two systems, choosing between them comes down

to the needs and requirements of the organisation and, to a lesser extent, the personal needs and preferences of those using it'. ¹⁷ However, open source log file analysis has the benefit of transparency, although a possible drawback with AWStats is its scalability. Although AWStats has been successfully tested with 10 GB log files, an operating system one hundred times the size of log files used for PCEL, this product may not be suitable for much larger sites.

A similar study was published in 2004, analysing the usage statistics for the National electronic Library of Infection (NeLI).¹⁸ Although comparisons are weakened by the fact that reported data for the NeLI study runs from January 2002 to June 2003, similarities and differences can be observed. In terms of web traffic, PCEL and NeLI are similar in magnitude: at the end

of the studies NeLI received 2866 page requests per month whereas PCEL received 10 306. Both PCEL and NeLI reported UK usage rates at around 33%. One marked difference was the percentage of NHS users: PCEL reported between 10% and 30% over the nine months of the study, whereas NeLI reported between 2% and 5%. Similarities exist in searching and browsing behaviour, and for requests per visit; however, PCEL would seem to be reaching more users and a higher percentage of NHS users than NeLI during this period. More up-to-date reports for NeLI usage can be found online, and report higher traffic for the period August 2003 to April 2004: 4500 page requests per month and 2000 unique visitors per month.¹⁹

This study suggests that improvements in the service provision of a digital library lead to increased usage. Although quantifying increases in popularity, the research does not qualify the reasons why more people are using the site more often. In attempting to answer this question we would like to avoid 'assessment by experts' and discover through a series of user questionnaires the qualitative aspects of PCEL to which users respond, both positive and negative.

Conclusions

This study shows that the PCEL represents an established UK-based primary care internet resource, and as such is a solid platform for further research. The resource has been online for over five years and continues to serve UK and NHS users; whilst usage has increased, use by UK and NHS users has not decreased. PCEL has become increasingly popular as defined by increasing numbers of visitors, visits and page requests during a planned period of improvement.

ACKNOWLEDGEMENTS

We are grateful to members of the Primary Care Informatics team at St George's and primary care professionals who have abstracted material for PCEL.

REFERENCES

- 1 Primary Care Electronic Library: www.pcel.info
- 2 National Library of Medicine. Medical Subject Headings (MeSH). www.nlm.nih.gov/mesh/meshhome.html
- 3 St George's University of London. Primary Care Informatics. www.gpinformatics.org
- 4 World Wide Web Consortium. RSS (Rich Site Summary). www.w3.org/WAI/highlights/about-rss.html#whatis
- de Lusignan S. The Doctor's Desk: one vision of how to deliver the future information and communication needs of general practice. *Journal of Informatics in Primary Care* 1998;6(1):19–22. www.primis.nottingham. ac.uk/informatics/may98/may7.htm

- 6 de Lusignan S, Brown A and Pritchard K. The Primary Care National electronic Library for Health (NeLH-PC): a pilot of information-centred knowledge management for primary care www.nelh-pc.nhs.uk. <u>Informatics in Primary Care</u> 2002;10:99–105.
- 7 Gray JA and de Lusignan S. National electronic Library for Health (NeLH). British Medical Journal 1999; 319:476–9.
- 8 Open Source Initiative (OSI). The Open Source Definition 2005. www.opensource.org/docs/definition.php
- 9 Wheatley M. The Myths of Open Source. *CIO Magazine* 2004. www.cio.com/archive/030104/open.html
- 10 Sagewerks. Website terminology 2005. sagewerks.com/ articles/website%20terminology
- 11 The Apache Software Foundation 2005: <u>www.apache.</u> org/licenses/
- 12 Netcraft. Web Server Survey Archives 2005. news. netcraft.com/archives/web_server_survey.html
- 13 Analog the most popular logfile analyser in the world: www.analog.cx/
- 14 AWStats official website: www.awstats.org/
- 15 Become.com: www.become.com
- 16 de Lusignan S, Pritchard K and Chan T. A knowledgemanagement model for clinical practice. *Journal of Postgraduate Medicine* 2002;48:297–303. www.jpgmonline. com
- Brown M. IIS vs. Apache. Looking Beyond the Rhetoric. ServerWatch 2003. www.serverwatch.com/tutorials/ article.php/10825_3074841_1
- Madle G, Kostkova P and Mani-Saada J. Evaluating the Online Activity and Searching Behaviour of Users of a Medical Digital Library. City University 2004. www.soi.city.ac.uk/~patty/HDL2004/HDL%202004%20 Workshop%20Abstracts.html
- 19 Report of the Evaluation of Access Logs of the NeLI August 2003 to April 2004. City University 2004. www.staff.city.ac.uk/~jane/WeblogsReport/Weblogs2004.htm

This paper was presented at the PHCSG Annual Conference 2005, Downing College, Cambridge, September 2005.

CONFLICTS OF INTEREST

None.

ADDRESS FOR CORRESPONDENCE

Simon de Lusignan Senior Lecturer, Primary Care Informatics Hunter Wing St George's – University of London London SW17 0RE

UK

Tel: +44 (0)208 725 5661 Fax: +44 (0)208 767 7697 Email: slusigna@sgul.ac.uk

Accepted September 2005