

Web Intelligence Analyses of Digital Libraries: A Case Study of the National electronic Library for Health (NeLH)

Alesia Zuccala¹, Mike Thelwall¹, Charles Oppenheim², Rajveen Dhiensa²

¹ School of Computing and Information Technology, University of Wolverhampton, Technology Centre (MI Building), Wulfruna Street, Wolverhampton, WV1 1SB, U.K. E-mail: A.Zuccala@wlv.ac.uk; M.Thelwall@wlv.ac.uk

² Department of Information Science, Loughborough University, Loughborough, Leicestershire, LE11 3TU, U.K., E-mail: C.Oppenheim@lboro.ac.uk; R.Dhiensa@lboro.ac.uk

Abstract

Purpose – To explore the use of *LexiURL* as a Web intelligence tool for collecting and analysing links to digital libraries, focusing specifically on the National electronic Library for Health (NeLH).

Design/methodology/approach – The Web intelligence techniques in this study; a combination of link analysis (web structure mining), web server log file analysis (web usage mining), and text analysis (web content mining), utilize the power of commercial search engines and draw upon the information science fields of bibliometrics and webometrics. *LexiURL* is a computer program designed to calculate summary statistics for lists of links or URLs. Its output is a series of standard reports, for example listing and counting all of the different domain names in the data.

Findings – Link data, when analysed together with user transaction log files (i.e., Web referring domains) can provide insights into who is using a digital library and when, and who could be using the digital library if they are “surfing” a particular part of the Web; in this case any site that is linked to or colinked with the NeLH. This study found that the NeLH was embedded in a multifaceted Web context, including many governmental, educational, commercial and organisational sites, with the most interesting being sites from the .edu domain, representing American Universities. Not many links directed to the NeLH were followed on September 25, 2005 (the date of the log file analysis and link extraction analysis), which means that users who access the digital library have been arriving at the site via only a few select links, bookmarks and search engine searches, or non-electronic sources.

Research limitations/implications – *LexiURL* uses the Yahoo! API for its link extraction, but the use of commercial search engine data has several limitations. First, no search engine covers the entire web and so all are likely to return incomplete results. This problem is exacerbated by the typical limitation of 1,000 results per query and for Google and Yahoo! and their automatic search services report only a fraction of the results known by the parent search engine. Hence for a large digital library, *LexiURL* could be expected to find perhaps only 10% or less of the links to the site. A second limitation is that the method by which each search engine finds pages is unknown, as is the method for ranking results.

Originality/value – A few studies focusing on digital library users have been carried out using log file analysis as a research tool. Log files focus on real-time user transactions; while *LexiURL* can be used to extract links and colinks associated with a digital library’s “organic” Web network. This Web network is often not recognized enough, and can be a valuable indication of where potential users are surfing, even if they have not yet specifically visited the NeLH site.

Keywords Digital Libraries, Web link Analysis, Log File Analysis, Web Intelligence,

Paper type Research paper

1. Introduction

The past decade has seen an enormous growth in the development of digital libraries, archives and repositories due to the Web's potential to provide easy access to information without geographic constraints. With the introduction of digital information services, there has been an increasing need for information professionals to develop new forms of evaluation. Once a digital library, archive or repository has been set up and its design optimised, stakeholders and managers need to know how useful it is, who is finding it useful, what it is being used for, and where it is not being used in instances when it might be particularly useful.

In this paper we present a complementary set of research techniques, which can be employed to help a digital library manager attain insight into when and how their Web information resources are being used, to identify opportunities for further use, and detect where special partnerships may be formed with "neighbouring" organisations (i.e., Web network allies). The techniques, a combination of link analysis (web structure mining); web server log file analysis (web usage mining); and text analysis (web content mining), utilize the power of commercial search engines and draw upon the information science fields of bibliometrics and webometrics. Details concerning each research technique are provided, including a case study and discussion of the findings resulting from a Web intelligence project carried out in 2005 for the National electronic Library for Health (NeLH).

2. Literature Review

In recent years, research concerning digital libraries has focused on questions of website design, information provision and information retrieval. Digital library research draws mainly from computer science (Witten, Boddie, Bainbridge, and McNab, 2000), information science (Fox and Urs, 2002) and library and information studies (Tuominen, Talja, & Savolainen, 2003).

Scholars have examined the usability of digital libraries both in terms of general information seeking and browsing (Blandford, Buchanin and Jones, 2004; McKay *et al.*, 2004; Borgman *et al.*, 2005), including a system's ability to facilitate a user's overall "information journey" (Adams and Blandford, 2005). Different types of user's will have varying needs and these needs are likely to change over time. A digital library must therefore be proactive: it cannot be a "passive warehouse of static information," but rather a system designed to "support users' overall information work in context" (Adams and Blandford, 2005, pp. 160-161). Browsing behaviour, which is often associated with less specific user needs, is also vital to the information seeking process. Browsing allows for serendipity and gives the information seeker an opportunity to re-think and re-evaluate an information need. McKay *et al.*'s (2004) research demonstrates how the usability of a digital library system is enhanced when an information seeker has the option of combining searching with browsing and can browse digital resources by more than one type of metadata.

The publication impact of digital library materials has been assessed (Kaplan and Nelson, 2000), as well as critical issues concerning preservation and sustainability (Tibbo, 2003; Ross and Headstrom, 2005). Society as a whole benefits when digital information is preserved both effectively and affordably. Research institutions have better opportunities for long-term cross-disciplinary collaboration, especially if collaborations depend on scientific data that is impossible to reproduce. Archives, museums and libraries can protect and conserve cultural memory, and in situations where accountability is crucial, hospitals, clinics and other public institutions are in a better position guarantee the authenticity and integrity of digital materials over time (Ross and Headstrom, 2005).

From a computer science perspective there has been a growing concern with online security (Atluri and Ray, 2004) and the enhancement of digital library visuals (Marks *et al.*, 2005). Digital libraries are situated in unique environments, thus environmental factors have also been examined for their impact on users' awareness, acceptance and use habits (Adams and Blandford, 2004). Often the social

contexts or domains surrounding a digital library are very different, for instance, clinical versus academic. Adams and Blandford (2004) demonstrate why it is important within academic contexts to market a digital library as a learning resource, but to show users within a clinical domain (e.g., a health care setting) how a digital library system can support evidence-based medicine with information that is “paramount” to decision making (p. 71).

Some digital libraries or archives, for example, the Internet Archive (www.archive.org) and the BBC collection of all web news reports (<http://news.bbc.co.uk>) exist or have been made possible because their content is already in digital form. In the UK and in other parts of the world there is also a movement towards preserving e-government websites. At the UK National Archives, e-government website preservation is a reaction to the fact that “the World Wide Web is increasingly becoming the principle means of interaction between Government, citizens and business” (The National Archives, 2005). The drive to digital is so strong, however, that numerous archives are now being created through the digitisation of physical objects in various ways. Examples include photographs of art and museum exhibits, World War II reconnaissance photographs (e.g., The Aerial Reconnaissance Archives, 2004) as well as more exotic forms such as three-dimensional virtual reality archaeological reconstructions.

Digital *repositories*, like digital libraries are collections assembled from digital objects. Heery and Anderson (2005) distinguish repositories from digital libraries in line with the following characteristics: 1) content is *deposited* in a repository (either by the content creator or third party); 2) the repository architecture manages content as well as metadata; 3) the repository offers a minimum set of basic services (e.g., put, get, search) and 4) the repository must be sustainable, trusted, well supported and well-managed (p.2). Specific types include *e-learning repositories*, *research data repositories*, and *institutional e-prints repositories*. The missions of each vary, in that *e-learning repositories* secure digitised learning resources, so that educators can make them available to students (e.g., online tutorial images or videos) or use them in their own work (e.g., lesson plans), while an *institutional e-prints repository*, localized within a university, contains items pertaining to the research output of its community members (e.g., preprints, journal articles, and conference papers). *Research data repositories*, such as The UK Data Archive (2005), support national or international studies through the provision of preserved digital data sets in the social sciences or sciences.

The primary rationale for creating a localized e-prints repository is that it will increase the visibility of a university’s research and perhaps enhance its prestige (Crow, 2002). Such a repository will not exist however, unless managers can motivate scholars to self-archive their work and help them recognize that doing so is to their benefit (e.g., Swan and Brown, 2004a,b). The issue of copyright is significant in an open access environment: authors are not necessarily aware of their legal right to distribute published materials online, nor have they signed such rights away. Fortunately more and more journals are giving authors the “green light” to self-archive (e.g., Gadd, Oppenheim and Proberts, 2003a,b,c). Once this practice becomes commonplace, research concerning how often e-prints are downloaded and cited by users will grow in situations where an institution and individual’s citation impact factor is paramount (e.g. Brody *et al.*, 2004; Harnad and Brody, 2004).

With certain types of institutional repositories consideration needs to be given to their benefit to users outside institutional boundaries, including users of other repository types. Greater forms of cooperation between digital libraries and repositories will be needed in the future, yet one of the most important challenges with interoperability is the development of facilitating software tools. Heery and Anderson’s (2005) assessment of the ecology of repositories state that there is at present “very little interoperability ... For example, e-print institutional repositories are unlikely to be linked to or interact with repositories for teaching and learning” (p.14). McLean and Lynch’s (2004) interoperability framework suggests that a lot of “work needs to be done involving standards, architectural modelling or interfaces (as opposed to cultural, organizational or practice questions in order to permit [information environments and learning environments] to co-exist and co-evolve more productively” (p. 1).

For digital library, archive and repository managers, the Web is not just a valuable access point or host for digitised information, but also an important resource for acquiring Web intelligence. Specific intelligence techniques, utilizing inlinks, outlinks, colinks, web text mining and log file data analyses play a role in understanding and evaluating user or potential user behaviours. Web intelligence is often associated with business intelligence, particularly in the areas of marketing (Lau *et al*, 2004) and performance assessment (Vaughan, 2004 a,b). Computing software systems are now being developed to solve problems related to knowledge management or to provide businesses with their own specialized intelligence agents (Marshall *et al.*, 2004). In the world of business we expect managers to take interest in the commercial Web, however, Web intelligence, with a focus on link data and link analyses is multifaceted and can be applied to many other organisations and institutions (Thelwall, 2004a; Tang and Thelwall, 2004b). To date, Web intelligence has not been discussed at length concerning digital libraries, archives, or repositories, thus our case study based on the National electronic Library for Health (NeLH) is designed to demonstrate both its technical feasibility and potential for generating useful information.

3. Web Intelligence Techniques

Three main sources of information can be used to evaluate a Web-based digital library: 1) link analysis for links associated with a site based on pages outside the site (i.e., Web structure mining); 2) Web server log file analysis for search engine queries or referrer pages (i.e., Web usage mining); and 3) text analysis for pages outside a site but associated with it in some way (i.e., Web content mining).

3.1 Direct links or inlink pages:

Links to a page are valued because, at an abstract level, each one represents an endorsement of the target page by the author of the source page. If pages on a Web site receive a lot of links from many places around the web, this means that the site as a whole is probably worth looking at. Also, each page that produces a link to a site will contain information about why a person or group might view that site as being worth a visit. Webometric researchers now refer to this as the “link motivation” or reason for linking (e.g., Kim, 2000; Thelwall, 2003; Bar-Ilan, 2005).

Even when links are not followed, and do not generate new visitors directly, they may still be helpful to improve the site’s search engine ranking. In this case, links are used as an indicator of target page importance (Brin and Page, 1998); a new computer science development drawn from citation analysis (Borgman and Furner, 2002). Whilst followed links to a Web site may be found by the target Web site owner from Web server log files, links that are not followed may need to be found through using advanced queries in search engines like Google or AltaVista. Such advanced queries would allow users to request pages linking to a given URL or site respectively. These facilities are even more useful when Web server logs are unavailable; however, it is important to note that the results are restricted to the coverage of the search engine.

3.2 Colinked pages:

In bibliometrics research, where there is a long tradition of analysing citations to extract information about bodies of literature (Borgman and Furner, 2002), co-citations have been established as a useful data type (Small, 1973). Two papers A and B are co-cited if there is a third paper C that references (cites) both of them. If papers A and B are frequently co-cited then this is a useful indicator that they have something in common (White and Griffith, 1982). The same may also be true of web pages or websites. Pages/sites A and B are co-inlinked if there is a page C that links to both (Björneborn and Ingwersen, 2004). If two sites or pages are highly co-inlinked then this may mean that they have something interesting in common. For a given website, the set of pages and sites with which it is highly co-linked would be worth investigating, for example to assess whether there is enough overlap or commonality of purpose to set up some type of formal collaboration.

3.3 Web server log file analysis / search queries:

Web servers typically create a log of all requests for information sent to them. Each log entry normally contains the URL of the request and the time and date of the request. Additionally the log may contain information about the user sending the request, such as their Internet location identifier (I.P. address or domain name), the type of browser or operating system they are using, and the URL of the page that they previously visited. Cookies may also be used to track individual users. This information is useful to website owners because it reveals basic facts about how often a site is visited and which pages are the most popular (Nicholas, Huntington, Lievesley, and Wasti, 2000). Computer scientists have exploited log files for a number of reasons, for example, to examine users' search motivations and search strategies (query strings) at OPACs (Online Public Access Catalogs) and digital libraries (Peters, 1993; Jones, Cunningham, McNab and Boddie, 2000) or to create website navigation aids (Wheeldon and Levene, 2003).

Web server log files can give two types of information about how a particular site is perceived and used (Thelwall, 2001), and it is these which are most useful for the current research. The conveyor of the relevant information is the URL of the previously visited page. When this page is on another site, this normally means that the other site contains a link that points to the server's site, and that the user has clicked on the link. Visiting the link source page should give some context about why the link was thought to be useful.

The second type of information embedded in some previously visited page URLs is search engine keyword searches. Most search engines embed the user's search terms in the URL of the results page, because this is one of two standard ways to store user submitted Web queries. This means that if the user clicks on a search result, then their search terms may find themselves in the log file of the target page's web server, via the URL of the search results page. Analysing the set of terms and phrases used in keyword searches that lead visitors to a site is a useful way of gaining insights into why visitors chose to come to the site and what they were hoping to find in it.

Web browsers do not always send the URL of the previously visited page, known as the 'HTTP Referrer' field, so this information is likely to be partial. Moreover, many technical issues render all log file data unreliable in various ways (Nicholas, Huntington, Lievesley, and Withey, 1999; Spink and Jensen, 2004) but log file analysis is still a good source of insights and user behaviour estimates.

4. Case Study

In this paper, we adopt a case study approach to analysing a single digital library: the National electronic Library for Health. Our general focus is to determine whether or not the web intelligence techniques that we outline above can give us useful information about a digital library's web network, in particular, information that can function as a complementary supplement to a log file analysis.

4.1 About the National electronic Library for Health (NeLH)

Due to the large size of the UK's National Health Service (NHS)¹, a strategy was formulated in 1998 to create a national Web-based library that would be accessible to a wide variety of healthcare professionals – i.e., doctors, nurses and other allied practitioners. The rationale for creating the NeLH was that many healthcare professionals could not always get to a library (in the physical sense), thus with a Web-based service, anyone with access to the Internet would be able to download and use any of the NeLH's accredited clinical reference material.

¹ The UK's National Health Service (as of 2005) employs approximately 1.2 million people. About 700,000 of these employees are potential users of the NeLH.

The NeLH was launched in November 2000 under the provision of the NHS Information Authority and was made available through the following URL: <http://www.nelh.nhs.uk/>. Today it is accessible to the public 24 hours a day, 7 days per week, and 365 days a year through any Web browser, and its mission statement is to “provide health care professionals with the best current knowledge and know-how to support health care related decisions” (National electronic Library for Health, 2005). Figure 1 below shows the NeLH’s entry page (see Figure 1).

The information content of the NeLH is comprised of three parts: 1) *knowledge-based information* (e.g., Clinical databases; MEDLINE; Full-text journals), 2) *know-how information* (e.g., The National Institute for Clinical Excellence/ NICE), and 3) *resources* (e.g., Specialist Libraries; Libraries by Profession; Medical Dictionaries). One of its most interesting features is the “Hitting the Headlines Project,” which allows clinicians and patients to learn more about the reliability of health-related news stories and the research evidence on which they are based. The NeLH has:

commissioned the Centre for Reviews and Dissemination (CRD) at the University of York, to assess the reliability of both the journalists' reporting of health stories and the research on which they are based. CRD staff provide a rapid assessment of the original research behind the news story and evaluate how accurately the journalists have reported the findings of the research. CRD produces summaries of news stories within 48 hours of their publication (NeLH, 2005).

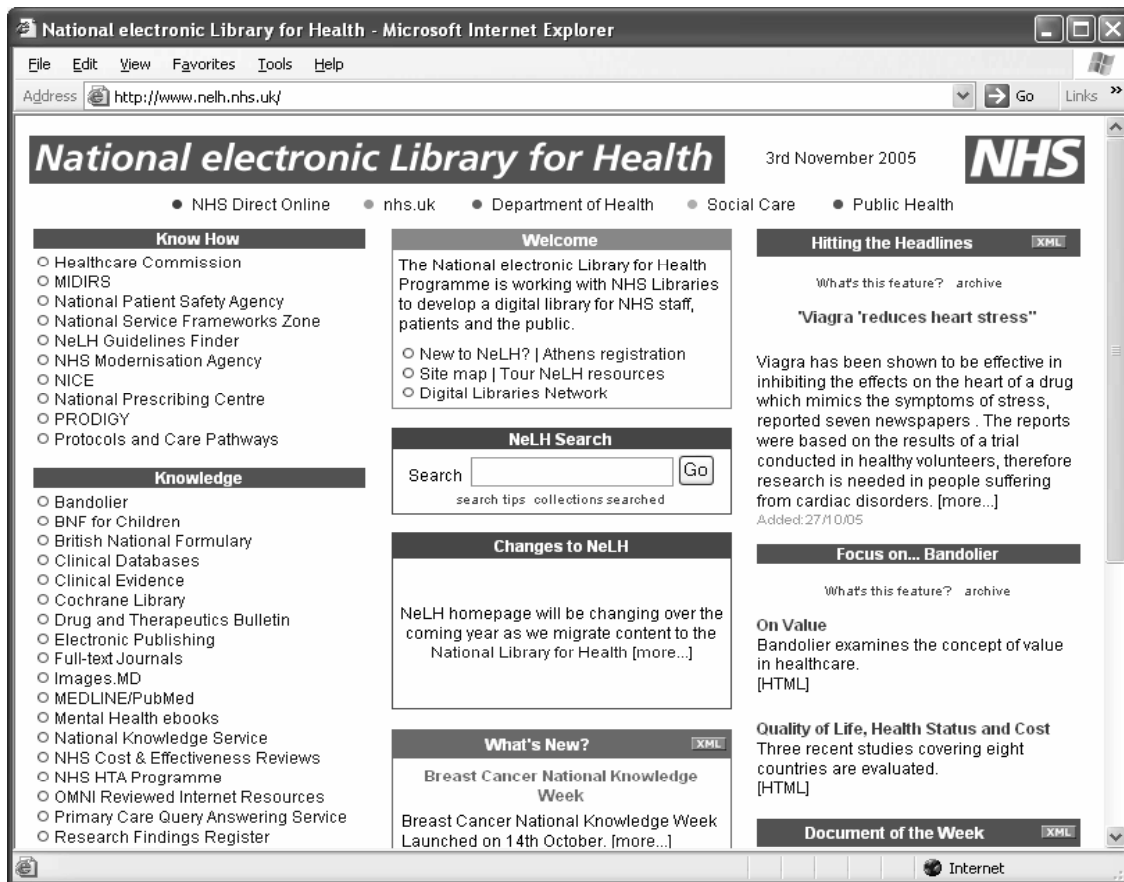


Figure 1. Entry page to the NeLH (2005).

Anyone using the NeLH can search through its databases from one convenient Pilot search portal, though some resources are password controlled. If a user is not familiar with the NeLH, online tour links have been set up at the web site (e.g., The NeLH Overview Tour) with, in some cases, built-in audio guides. This shows the degree to which the owners have taken measures to ensure that the site is used as widely and effectively as possible.

4.2 Meeting with the NeLH Service Delivery Coordinator and Case Study Objectives

On October 4, 2005, two members of our research group held a meeting with the Service Delivery Coordinator of the National electronic Library for Health (NeLH) to investigate this particular digital library's service goals and present a descriptive report containing newly retrieved link and colink data. The data were collected on September 20th, 2005 using a special software tool – i.e., *LexiURL* (2005) – which extracts data from the Yahoo! search engine. During this meeting, questions associated with the Web link report were discussed (e.g., *What are your initial thoughts about the link data? Is there anything about this report that surprises you? How often would you like to receive a report like this?*), and one of the main outcomes was the following set of questions that we developed for this case study:

- 1) What do the direct links to the NeLH reveal about its general visibility on the Web?
- 2) What can we learn from comparing the NeLH's *WebTrends* log file data with the link statistics collected using *LexiURL*?
- 3) Are there any “hidden” websites associated with the NeLH that might be of interest to the service manager?
- 4) Can comparative link statistics concerning a similar digital library on the Web – i.e., the National Library of Medicine (U.S.) – provide the NeLH with valuable strategic information?

5. Link Statistics and Web Transaction Log File Data

5.1 *LexiURL* link extraction

With the *LexiURL* link extraction software, a search was carried out with Yahoo! for links to the National Electronic Library for Health (NeLH). The following query does not just match links to the NeLH home page, but links to any page within the NeLH domain, and excludes internal site links (site self-links):

linkdomain:www.nelh.nhs.uk NOT site:www.nelh.nhs.uk

The query reported approximately 39,000 matching pages. We retrieved the first 999 URLs from Yahoo! (the maximum allowed) and 887 pages were successfully downloaded and verified as containing a link to one of the pages on the NeLH site. Table IA presents an alphabetically sorted list of 50 downloaded pages (see Appendix). Note that no search engine covers the entire web and so all are likely to return incomplete results. This problem is exacerbated by the typical limitation of 1,000 results per query and for Google and Yahoo! their automatic search services report only a fraction of the results known by the parent search engine. Hence for a large digital library, *LexiURL* could be expected to find perhaps only 10% or less of the links to the site. Also, the method by which each search engine finds pages is unknown, as is the method for ranking results, except in broad detail (Chakrabarti, 2003; Levine, 2005).

The 887 pages were analysed to determine the number of pages in each site that link to the NeLH and the number of different domains from that site containing at least one page that links the NeLH. In this context Web sites are identified with domain name endings; including the second or top-level

domain and one additional domain name segment. Table I presents a partial list of the site domain numbers and page numbers sorted in descending order based on the highest page count (see Table I).

Table I. A list of site pages (N=20) and domains that link to the NeLH.

Site	Domains (Number of different domains in the site that contain at least one page that links to the web site www.nelh.nhs.uk)	Pages (Number of pages in the site that link to the web site www.nelh.nhs.uk)
hyscience.com	1	350
omnimedicalsearch.com	1	121
blogspot.com	10	70
tabebak.com	1	39
gfmer.ch	1	17
freepint.com	1	12
cebm.net	1	5
edu.pe	1	5
wikipedia.org	1	4
obels.org	1	4
factbites.com	1	4
earlham.edu	1	3
typepad.com	2	3
virtualchase.com	1	3
weblogs.com	1	3
miami.edu	2	3
bmjournals.com	1	3
kmblogs.com	1	3
uni-lj.si	1	3
gillesenvrac.ca	1	3

In some cases, many links came from one Web site. For example, 350 pages linking to the NeLH came directly from hyscience.com. In some cases the links came from one site with different domain names. For instance, 70 different pages from blogspot.com were linked to the NeLH, but each inlink page was from one of 10 different domains, each being a separate blog. Sites with many linking pages could either have replicated links (e.g., in a navigation bar throughout the site) or could find the NeLH extremely useful. Investigations of individual pages are needed to identify which applies in each case.

1. **quackfiles**.blogspot.com (60 pages)
2. **zillman**.blogspot.com (2 pages)
3. **lupusnewslog**.blogspot.com (1 page)
4. **healthcareresources**.blogspot.com (1 page)
5. **dlnet**.blogspot.com (1 page)
6. **socialinformatics**.blogspot.com (1 page)
7. **liberatingtheliterature**.blogspot.com (1 page)
8. **kegliography**.blogspot.com (1 page)
9. **poorfrank**.blogspot.com (1 page)
10. **nsi**.blogspot.com (1 page)

Figure 2 presents a ranked list of second or top level (SLD) domains linking to www.nelh.nhs.uk and the number of sites associated with each SLD (see Figure 2). Note that most of the inlinks were from the .com, .org, .edu and .net domains. The .ca domain shows that Canada has a few links to the NeLH site, as well as Belgium (.be), The Netherlands (.nl) and Denmark (.dk).

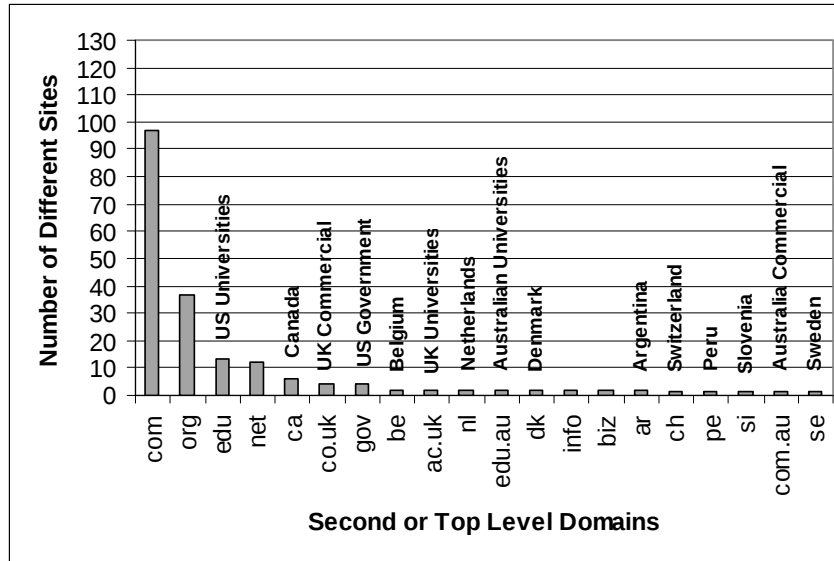


Figure 2. Second or Top Level Domains (SLDs) and Number of Site Inlinks to the NeLH.

5.2 Web usage log files

WebTrends log files for the NeLH were collected on September 20th, 2005: the same day that the link data were extracted from the Web using *LexiURL*. In Table II, we present a list of referring domains for the NeLH (see Table II). A referring domains report indicates which domains on the Web are sending the most traffic to the NeLH site. Table III shows what percentage of the NeLH transaction files came from a Web page (instead of another source) and Table IV presents a sample of the most frequent search engine phrases that visitors used before arriving at the NeLH homepage (see Tables III and IV).

Table II. NeLH referring domain report.

Referring Domain	Occurrences
No Referrer	31012
http://libraries.nelh.nhs.uk	3235
http://hermes	3117
http://google.co.uk	1174
http://google.com	457
http://search.library.nhs.uk	333
http://intranet	314
http://nhstdirect.nhs.uk:8080	304
http://rms.nelh.nhs.uk	296
http://le.ac.uk	275
http://wkp.nhs.uk	233
http://216.239.59.104	194

http://search.msn.co.uk	164
http://johnsquirelibrary.org.uk	136
http://images.google.com	128
http://base-library.nhs.uk	107
http://cochrane.org	98
http://nww.berkshire.nhs.uk	92
http://hop.man.ac.uk	90
http://search.yahoo.com	90
http://ferrybridge	83
http://nhsdirect.nhs.uk	77
http://nmht	73
http://uk.search.yahoo.com	72
http://google.ca	70
http://aditus.nhs.uk	57
http://patient.co.uk	52
http://webserver	50
http://p-jones.demon.co.uk	49
http://radio.userland.com	48

Table III. Number and percentage of hits associated with a specific NeLH referrer origin.

Referrer Origin	Hits	Percentage
No Referrer	30895	67.14%
Web page (http://)	15108	32.83%
Other	16	0.03%

Table IV. Most frequent NeLH user search phrases.

Search Phrase	Search Engine	Hits
nelh	Google	203
nhs	Google	55
nelh	Microsoft Network	32
NHS	Google	32
national electronic library for health	Google	26
cochrane library	Google	24
NeLH	Google	23
nhs library	Google	19
clinical audit	Google	18
cochrane	Google	18
NeLH	Google	16

5.3 *LexiURL* colink extraction and mapping

To develop a colink map of the NeLH site and other sites on the web, we started with the original *LexiURL* link extraction of 887 inlink pages and filtered them down to the site level. We then located the top 49 sites colinked most frequently with the NeLH site and selected them for a Web colink analysis (see Table V).

The Web colink analysis is based on a raw data matrix, converted into a matrix of Pearson's *r* correlation coefficients and used as input into SPSS-12 for a multidimensional scaling (ALSCAL) routine. Table IIA, located in the Appendix presents the raw data and summary statistics (see Appendix). Figures 3 and 4 illustrate the resulting colink maps. Figure 3 shows the NeLH's site nodal position relative to 49 other sites on the Web and Figure 4 highlights specific relationships between the nodes. Previous research using Web colink data has shown that colink clusters can be thematically elusive and difficult to label if there is a complex interplay of underlying link motivations (Zuccala, 2006). A full SPSS-12 CLUSTER routine was not used in this analysis; however with some colink studies, which use data for confirmatory rather than exploratory purposes, CLUSTER routines are more applicable (e.g., Ortega-Priego and Aguillo, 2005; Vaughan and You, 2005).

5.4 Strategic Comparison with the United States' National Library of Medicine (NLM)

In addition to the NeLH, a second *LexiURL* link extraction was carried out for the United States National Library of Medicine (NLM). The National Library of Medicine (NLM) and the National electronic Library for Health (NeLH) are comparable services; however, the NLM has been publicly available on the Web for a much longer period and exists also as a physical space at the campus of the National Institutes of Health in Bethesda, Maryland. U.S. The digitalized version of the NLM was first launched in October of 1993 and has been marketed to Web users as "the world's largest medical library" (United States National Library of Medicine, 2005). Several of the information resources offered by the NLM, for instance, access to MEDLINE, PubMed and full-text journals, are similar to those offered at the NeLH. The NeLH however, places more emphasis on providing information to healthcare professionals, while the NLM is recognized as a key resource for professionals, scientific researchers and members of the general public, including patients and their families. Another difference between the two digital libraries is that the NLM houses a special digital collection on the History of Medicine, whereas the NeLH does not. The NLM is also currently the world's most valuable resource for researchers and clinical professionals interested in data associated with the Human Genome Project.

Again, with the *LexiURL* link extraction software, a search was carried out with Yahoo! for links to the National Library of Medicine (NLM), omitting all NLM site self-links. Our query did not just match links to the NLM home page, but links to any page within the NLM domain:

linkdomain:www.nlm.nih.gov NOT site:www.nlm.nih.gov

The query result was approximately 1,350,000 URLs of matching pages. We retrieved the first 999 URLs from Yahoo! and 597 pages were successfully downloaded and verified as having a link to a page on the NLM site. Table VI presents a partial list of individual site domain numbers and page numbers sorted in descending order based on the highest page count (see Table VI). Figure 5 shows the percentage of site inlinks shared between the National electronic Library for Health (NeLH) and the National Library for Medicine (NLM). Figure 6 presents the top ranking list of second or top level (SLD) domains linking to www.nlm.nih.gov and the number of sites associated with each SLD.

A colink map was also constructed based on the top 49 sites most frequently colinked with the National Library for Medicine (NLM) (see list in Table VII). Table IIA in the Appendix shows the partial raw data matrix and summary statistics (see Appendix). Figures 7 and 8, based on the same mapping procedure used for the NeLH colinks, are also shown below.

Table V. Top 49 sites colinked most frequently with the NeLH (N=50 sites inclusive). Sites with a direct link to the NeLH (including the NeLH) are highlighted in bold font.

about.com	About.com Internet Guide
ahcpr.gov	Agency For Health Care Policy and Research
ama-assn.org	American Medical Association - Helping Doctors Help Patients
aol.com	AOL: America Online
bbc.co.uk	BBC News
berkeley.edu	University of Berkeley California
blogspot.com	Blogger
bmj.com	Electronic BJM: British Medical Journal
bmjjournals.com	BMJ Journals Online: Essential Reading For Medical Specialists
cdc.gov	US Dept of Health & Human Services: Centers for Disease Control and Prevention
cnn.com	CNN News
columbia.edu	Columbia University
cornell.edu	Cornell University
dhhs.gov	United States Department of Health & Human Services
emedicine.com	Emedicine: Instant Access to the Minds of Medicine
emory.edu	Emory University, Atlanta, Georgia
fda.gov	US Food and Drug Administration
gc.ca	Government of Canada
geocities.com	Geocities Free Web Hosting
guideline.gov	National Guideline Clearinghouse: Public Resource for Clinical Practice Guidelines
harvard.edu	Harvard University
healthfinder.gov	US Department of Health & Human Services: Healthfinder
hon.ch	Health on the Net Foundation (Geneva, Switzerland)
intelihealth.com	AETNA Intelihealth
ki.se	Karolinska Institute Sweden
loc.gov	Library of Congress
mcmaster.ca	McMaster University Canada
medscape.com	Medscape WebMD Medical Information
merck.com	Merck & Co. Pharmaceuticals
msn.com	Microsoft MSN
nelh.nhs.uk	National Electronic Library for Health
nih.gov	US Department of Health & Human Services: National Institutes of Health
ox.ac.uk	University of Oxford
shef.ac.uk	University of Sheffield
stanford.edu	Stanford University
tripod.com	Tripod: Create Your Own Website
ucla.edu	University of California Los Angeles
ucsf.edu	University of California, San Francisco
uiowa.edu	University of Iowa
uiuc.edu	University of Illinois at Urbana-Champaign
umich.edu	University of Michigan
upenn.edu	University of Pennsylvania
usda.gov	United States Department of Agriculture
vh.org	Virtual Hospital: A Digital Library of Health Information (University of Iowa)
washington.edu	University of Washington
webmd.com	WebMD Inc.
who.int	World Health Organization
wisc.edu	University of Wisconsin-Madison
yale.edu	Yale University
york.ac.uk	University of York

Table VI. A list of site pages (N=20) and domains that link to the NLM.

Site	Domains (Number of different domains in the site that contain at least one page that links to the web site www.nlm.nih.gov)	Pages (Number of pages in the site that link to the web site www.nlm.nih.gov)
us.com	1	46
nih.gov	25	40
uiowa.edu	2	27
yahoo.com	1	10
fda.gov	3	8
resourceshelf.com	1	7
refdesk.com	1	7
cdc.gov	2	7
ki.se	1	6
blogspot.com	6	6
coreynahman.com	1	6
pitt.edu	2	6
wisc.edu	2	6
nihseniorhealth.gov	1	5
nnlm.gov	1	4
tmc.edu	1	4
wustl.edu	1	3
eskimo.com	1	3
ruki.org	1	3
khake.com	1	3

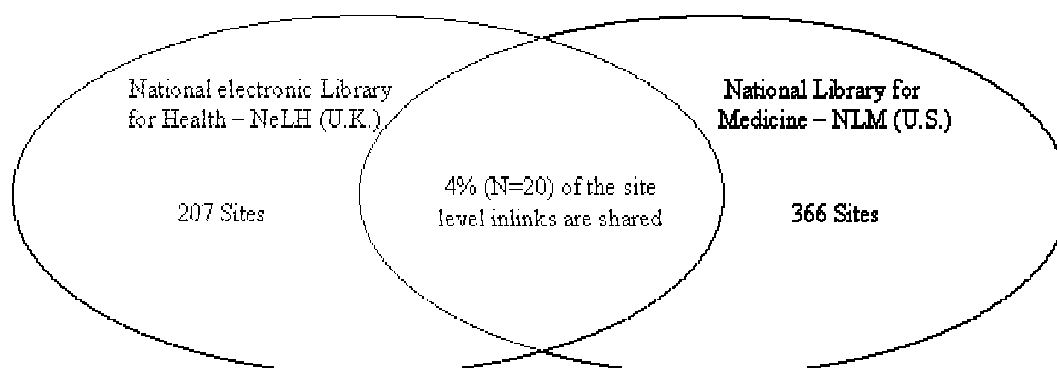


Figure 5. Percentage of site inlinks shared between the NeLH and NLM.

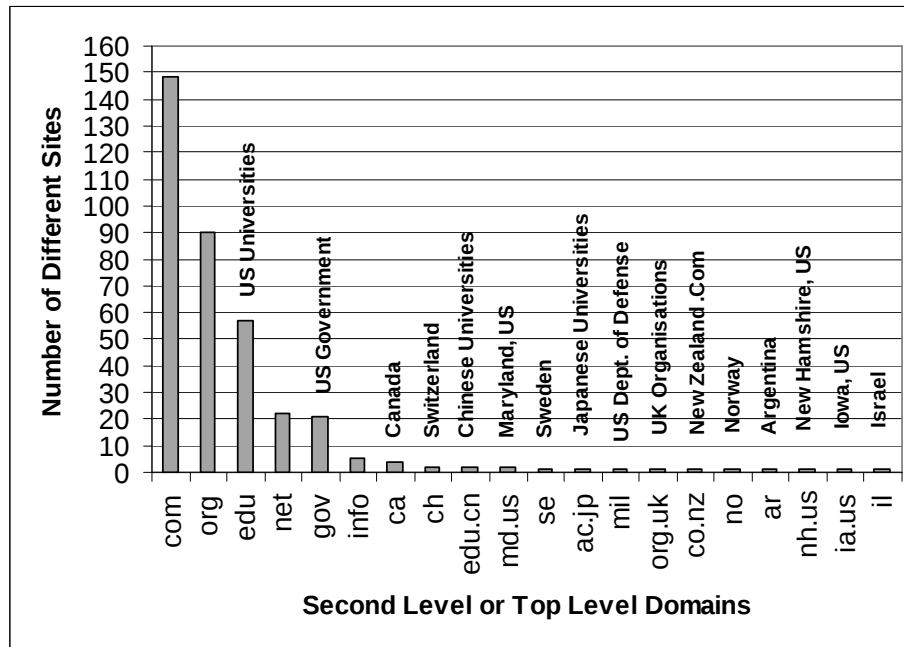


Figure 6. Second or Top Level Domains (SLDs) and Number of Site Inlinks to the NLM.

6. Web Intelligence Analysis of the Link and Colink Data

As a research concept, “Web intelligence” can mean different things. In the field of computer science, for example, “Web intelligence” is seen as “a new direction for scientific research that explores the fundamental roles as well as practical impacts of artificial intelligence and advanced information technology for the next generation of Web-empowered system, services and environments” (Web Intelligence Newsletter, 2003; Zhong, Liu, and Yao, 2003). In this study, Web intelligence assumes a user-centred focus: we defined it as the process by which link data is collected, organized, and analysed for insights, which may be used by managers for strategic decision making purposes.

Sometimes a Web hyperlink is created for a basic technical reason rather than for any real communication function, but often a directed link has a purpose. A review of the link’s context can help identify its purpose (Thelwall, 2004). Bar-Ilan’s (2005) advice for reviewing direct links is to focus on the characteristics of the source page, that is, the relationship between the source of the link and the link target. Consideration also needs to be given to where it is located on the source page, how the link creator discusses it, and whether or not there is any link tone (e.g., *Is it positive, neutral or negative?*). Link counts can be difficult to work with, particularly if they are extracted in bulk; thus it is often useful to order the data by sorting URLs alphabetically by page, site or domain level. Ordering the links allows the investigator to determine which ones stand out as being unexpected or interesting, and worth visiting for a contextual review.

Table VII. Top 49 sites colinked most frequently with the NLM (N=50 sites inclusive). Sites with a direct link to the NLM (including the NLM) are highlighted in bold font.

about.com	About.com Internet Guide
ama-assn.org	American Medical Association - Helping Doctors Help Patients
aol.com	AOL: America Online
arizona.edu	University of Arizona
bbc.co.uk	British Broadcasting Corporation
berkeley.edu	University of Berkeley California
cdc.gov	US Dept. of Health & Human Services: Centres for Disease Control and Prevention
cnn.com	CNN News
columbia.edu	Columbia University
cornell.edu	Cornell University
dhhs.gov	United States Department of Health & Human Services
duke.edu	Duke University (Durham, North Carolina)
emory.edu	Emory University, Atlanta, Georgia
fda.gov	US Food and Drug Administration
geocities.com	Geocities Free Web Hosting
harvard.edu	Harvard University
hhs.gov	United States Department of Health & Human Services
indiana.edu	Indiana University
jhu.edu	Johns Hopkins University
loc.gov	Library of Congress
medscape.com	Medscape WebMD Medical Information
merck.com	Merck & Co. Pharmaceuticals
mit.edu	Massachusetts Institute of Technology
nasa.gov	National Aeronautics and Space Administration
nlm.nih.gov	National Library for Medicine
nytimes.com	The New York Times
nyu.edu	New York University
ohio-state.edu	Ohio State University
pbs.org	PBS Home (Public Broadcasting Station)
pitt.edu	University of Pittsburgh
stanford.edu	Stanford University
tripod.com	Tripod: Create Your Own Website
uchicago.edu	University of Chicago
ucla.edu	University of California at Los Angeles
ucsf.edu	University of California, San Francisco
ufl.edu	University of Florida
uiowa.edu	University of Iowa
uiuc.edu	University of Illinois at Urbana-Champaign
umich.edu	University of Michigan
umn.edu	University of Minnesota
unc.edu	University of North Carolina
upenn.edu	University of Pennsylvania
usda.gov	United States Department of Agriculture
utexas.edu	University of Texas at Austin
virginia.edu	University of Virginia
washington.edu	University of Washington
who.int	World Health Organization
wisc.edu	University of Wisconsin-Madison
wustl.edu	Washington University in St. Louis
yale.edu	Yale University

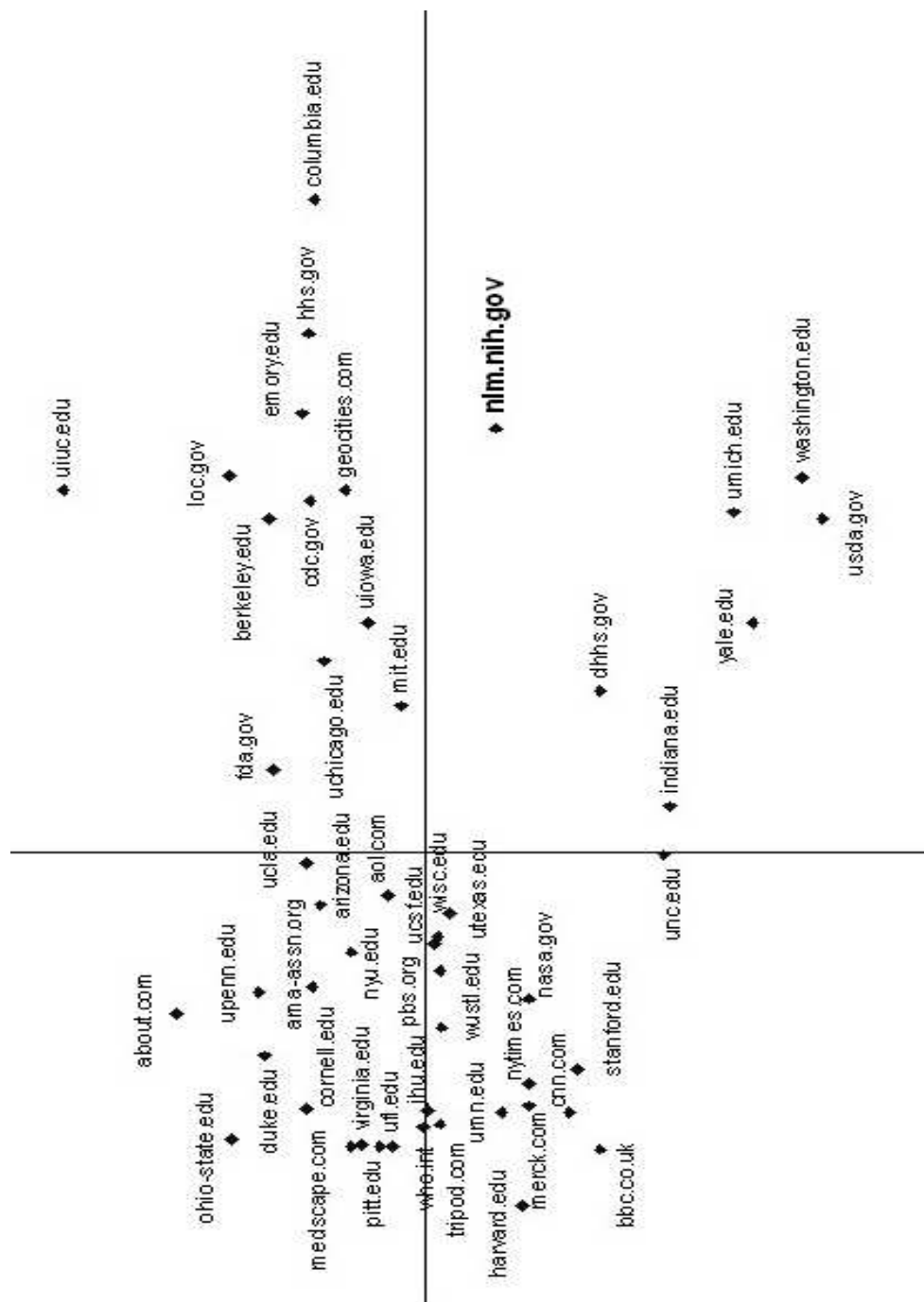


Figure 7. Top 49 web sites colinked with the NLM.
2D Scaling Stress Value = .146 and Proportion of Variance of the Scaled Data (RSQ) = .914

6.1 What do direct links to the NeLH reveal about its general visibility on the Web?

Amongst all of the links directed to the NeLH, the high frequency of .com sites was expected. Brief visits to the .com sites turned up a few search engines, website hosting services, and a variety of healthcare sites based on the following topics: *Asperger's Syndrome*; *Amniocentesis*; *Beauty, Dieting, Nutrition, Weight-loss and Anti-Aging products*; *Birth, Postpartum & Breastfeeding*; *Bladder Cancer*; *Cardiovascular Disorders*; *Care of the Elderly*; *Cervical Cancer*; *Dyslexia*; *Downs Syndrome*; *Hormone Replacement Therapy*; *Lupus*; *Medical Malpractice*; *Midwifery*; *Occupational Therapy Reviews and Trials*; *Palmtop Computing Resources for Health Care Professionals*; *Pregnancy, Prostate Cancer*; *Second-hand Smoke (Effects of)*; *Social Network Analysis*; *Vaccinations*. Most of the topics relate to the healthcare issues that doctors, nurses and other clinical specialists in the UK might encounter in their everyday practice; hence the NeLH seems to be attracting links from websites based on relevant concerns.

In addition to pages from the .com domain, we visited several other pages (i.e., from .net, .edu, .org etc.) and found links to the NeLH, which could be classified as “acknowledgement links” (Thelwall, 2003). Website developers are acknowledging the NeLH's presence on the Web, that is, directing users to its home page, but few are referring to its content areas. Some of the websites that we visited demonstrated a specific interest in the NeLH's *Specialist Library on Knowledge Management* (kmblogs.com; elearnspace.org; zillman.blogspot.com), the *Cancer Specialist Library* (cancerlibrary.net), the *Hitting the Headlines* page (gatewaytobeauty.com; bmjjournals.com), the *Specialist Library on Screening* (gfmer.ch; en.wikipedia.org) and the *Specialist Library - Guidelines Finder* (info.ki.se; library.adelaide.edu.au).

A selection of weblogs link to the NeLH, but again, most seem to be hosting acknowledgement links. Blog spaces allow authors to post discussions or debates on the Web and use (link to) recognized information sources to support their arguments. Some blogs have been criticized however for not being effective at achieving this purpose (e.g., Cronin, 2005). Two blogs that we visited, namely *Confessions of a Quackbuster* and *HyScience*, have hosted discussions on different healthcare topics (e.g., the link between a measles, mumps & rubella vaccination and autism; Cardiovascular risk and COX-2 inhibition in rheumatological practice), yet neither has commented on the NeLH's usefulness or referred to its content. A link to the NeLH was found at the *Quackbuster* blogspot under a navigation list titled “Other Great Websites.” At the *HyScience* blog we found another similar navigation list titled “Medical Blogs” where the authors link to the NeLH's *Clinical Evidence* search page. *Lupusnewslog*, another health-related blogspot, has included at least one contextual reference to the NeLH's *Hitting the Headlines* assessment page.

A few international healthcare organisations, institutes and universities link to the NeLH, for example, *The Karolinska Institute* (Sweden), *The Prince Leopold Institute of Tropical Medicine* (Belgium), and the *McGill University Health Centre-Nursing Research* (Canada), yet we were surprised to find only one hospital in our extraction list: the *Ball Memorial Hospital Library* (Muncie, Indiana, U.S., ballhospital.org/library). A brief Google search was carried out to locate hospitals in the UK with websites, and our search led us to the *NHS England Home* “All hospitals” index. The *Addenbrooke's Hospital* (addenbrookes.org.uk) in Cambridge hosts a link to the NeLH, but this did not turn up in our LexiURL extraction list. At the *Birmingham Women's Hospital* (bwhct.nhs.uk) we found another link to the NeLH *Specialist Library - Women's Health*. Not all hospitals across the UK manage a website; therefore it would be interesting to learn more about those that do, and why or why they do not link to the NeLH. The NeLH was created to make all of the National Health Service's accredited information available to clinicians online: a hospital website link can serve as a very useful entry point.

Less than 1% of the links directed to the NeLH were from personal home pages; however, previous research has shown that homepage links to digital libraries are rare (e.g., Beaulieu, 2005). Academics often create websites, but the motivation for this is much stronger than it is for healthcare professionals. Academics need to advertise their scholarly activities, and in doing so, some point to resources that interest them or that they use regularly for their research. Healthcare professionals, on

the other hand, work with people in hospitals, local surgeries or clinics. A personal website is not a tool that they need to facilitate their work, but if they do create one, it can be a useful entry point to the NeLH. Possibly also, personal Web publishing is relatively undeveloped and not resourced in the NHS. With our LexiURL software we retrieved one link to the NeLH from the links page of a Dr Abhishek Arora (<http://homepage.mac.com/srivastava/abhishek/page1.html>). Dr. Arora is a physician of Indian origin working in the United Kingdom on a surgical training scheme. He may or may not be a regular user of the NeLH but his link to this digital library is a good indicator of probable use. We also retrieved a link from the website of the late Professor Dr. Anne Clyde, at the University of Iceland (http://www.hi.is/~anne/online_services.html). Dr. Clyde taught courses on “The Internet for Research” and has written various articles concerning library weblogs and school library websites. Her research interests were not relevant to medicine; however, she was an information scientist with some expertise in evaluating internet resources. At her website, Dr. Clyde included a simple acknowledgement link to the NeLH, with no written evaluation, although she may have evaluated this resource in the past with the students in one of her courses.

6.2 What can we learn from comparing the NeLH’s *WebTrends* log file data with the link statistics collected using *LexiURL*?

Transaction logs function as a good source of usage information. Source URLs, keywords, IP Addresses – which can be obtained through a log file analyser (e.g., *WebTrends*) – can signify *where* users are coming from and *why* they might be visiting a particular site. *LexiURL* provides information that is different from a log file analyser because it harvests links from the Web as they exist “organically” in a small Web network. This small network includes inlinks and colinks associated with a target site, which in this case is the NeLH. Such links signal potential gateways or entry points for users. It is important to know where users are coming from, and what links they are following to access an online resource, but it can also be important to find out where users *could* be coming from on the Web, particularly if links exist, which are not being picked up by a transaction log.

When we compared the *WebTrends* log file data with the *LexiURL* link data we found almost no overlap between the two data sets. We combined the sets together to a total of 326 individual domain URLs, and as we scanned through them, we noticed that the log file URLs differed greatly from the *LexiURL* URLs, with the exception of four domains: 1) www.cebm.net, 2) www.cochrane.org, 3) www.guideline.gov, and 4) www.omnimedicalsearch.com. The *LexiURL* software extracted the four domains above as recognized links to the NeLH, while *WebTrends* recorded (on the same date) that users had actually followed them as an entry point to the digital library. Our comparison showed us which links were followed on September 20th, 2005, including a vast majority of links that were not followed (or at least not recorded), but could be in the future. On any given date, we expect that only a tiny majority of links to the NeLH will be followed. Some link contexts serve as convenient entry points (e.g., www.library.nhs.uk), while others are much less noticeable (e.g., lonestar9.pitas.com). Also, our *LexiURL* sampling system does not guarantee that all known links will be retrieved because it can harvest only a fraction of those highly ranked by Yahoo!, and Yahoo! itself does not cover the whole Web.

WebTrends provides a list of keywords and phrases associated with individual user searches, so we were able to determine the extent to which the NeLH has already become known. Table IV indicates that users have a habit of searching for this digital library by name (see Table IV); hence this is one of the reasons why google.co.uk and google.com came up as top referrer URLs (domains) in the log file. The *WebTrends* search phrases are also helpful for understanding why many of the *LexiURL* links to the NeLH were not followed on September 20th, 2005.

6.3 Are there any “hidden” websites associated with the NeLH that might be of interest to the service manager?

The word “hidden” is used to emphasize unrecognized websites associated with the NeLH, which possess either a direct link to this digital library or a colink in the same small Web network.

Sometimes direct links are “hidden” because users rarely follow them and they are not logged as regular access points. With the *LexiURL* link retrieval software it is possible to gather a sample of “hidden” URLs for review purposes to find out why they may or may not be visited regularly. For instance, in section 5.2 we note that *LexiURL* retrieved a few blogs on the Web with a direct link to the NeLH. Such links represent an important form of exposure for the NeLH, but within the context of a “noisy” blog space, visitors might not be enticed to follow an acknowledgement link because it is not situated within a discussion entry. Do general users need to see a clear reference to the NeLH’s content in order to become interested in what it has to offer?

We have identified some “hidden” hospital links to the NeLH, but with this discovery we recognize that the *LexiURL* software did not retrieve them through the Yahoo! ranking system. *LexiURL* does not have the capacity to harvest all known links, yet the sampling technique can still be informative. When an expected link doesn’t appear in a retrieval list, we know that a special target search should be carried out on the Web to learn more about its absence (Note: as we did by searching the *NHS England Home* “All hospitals” index to find hospital sites linking to the NeLH).

Other “hidden” websites may be identified through a colink mapping of the NeLH’s Web network. Figures 3 and 4 both show the top 49 websites colinked most frequently with the NeLH. Twelve of the websites possess both a direct link and colink to the NeLH (see Table V), but for this study we are interested in sites that are colinked, but NOT directly linked to the NeLH. Colinks represent “potential users” or at the very least, sites where potential users might be “surfing” the Web.

Note from Figure 4 that there is a fairly even distribution of colinked sites from the government, education, commercial, and organisational Web domains. Some of the commercial sites fit logically within the network structure. The *bmjournals.com* colink and direct link was expected because it is one of the NeLH’s content providers. A few website host spaces and search engines (e.g., *tripod.com*, *geocities.com* and *aol.com*) also appear on the map, but we expect these large, multifaceted commercial sites to appear, because this is where millions of users are active on the Web. Some important government sites are logical to this network because they are relevant to the medical/healthcare theme, for instance, the National Institutes of Health, U.S. (www.nih.gov) and the United States Department of Health and Human Services Centers for Disease Control and Prevention (www.cdc.gov). Of interest is the fact that these government sites are mainly American.

We can also see from Figures 3 and 4 that the NeLH has been colinked frequently with the BBC (bbc.co.uk). To determine whether or not the colink was motivated by the NeLH’s *Hitting the Headlines* resource, we examined a few link source pages. The Library page at the Prince Leopold Tropical Institute of Medicine colinked the nelh.nhs.uk with the bbc.co.uk/worldservice but each link was placed under different headings: “International Health, Reproductive Health, Population Issues, Development,” versus “News and Regional Information.” Also, *Tabebak.com*, an Arabic Web resource for the education of healthcare professionals in Egypt, had made repeated colinks between the nelh.nhs.uk and the bbc.co.uk, but these were created because the BBC hosts a special page written in Arabic, for Arabic news: bbc.co.uk/hi/arabic/news.

If we focus on all of the educational sites in the NeLH colink network (see Figures 3 and 4), most tend to be American (e.g., *Stanford*, *UCLA*, *Harvard*, *University of Wisconsin-Madison*, *Cornell*, *Columbia*, *Berkeley*, *University of Michigan*). Clearly, the important sites “hidden” from this network are the universities or educational institutes from the United Kingdom, with the exception of the *University of Sheffield*; *York University*, and the *University of Oxford*. In one sense it is impressive that the NeLH is part of an international Web network, including U.S. universities with prestigious medical schools, but on the other hand, the degree of American representation is puzzling. Is this a case of American sites and site linkages dominating the Web? Are UK sites not providing enough link support for their online comrades?

A follow-up direct link search using AltaVista [linkdomain:nelh.nhs.uk AND host:ac.uk] showed us that plenty of UK universities were and are linking to the NeLH (e.g., pages from the *University of*

Exeter; University College London; University of Kent; University of Sheffield; Liverpool John Moores University; University of Leicester; Lancaster University; Birbeck University of London, etc.). The LexiURL software had only retrieved direct links from the following universities: University of Newcastle upon Tyne and the University of Glasgow. Why more academic/educational sites did not turn up in our initial search, it is difficult to say. Again, our results are most likely based on the limitations of LexiURL as it harvests ac.uk links in conjunction with the Yahoo! ranking system. Academic sites in the UK are therefore not very highly ranked in comparison to many other sites.

6.4 Can comparative link statistics concerning a similar digital library on the Web – i.e., the National Library of Medicine (U.S.) – provide the NeLH with valuable strategic information?

Links to the National Library of Medicine (NLM) from the .com and .org domains were associated with a different set of healthcare topics than those associated with the NeLH: *Aids-HIV; Brain Aneurysm; Cannabis News; Celiac Disease; Chronic Pelvic Pain or Chronic Prostatitis/Cystitis; Cosmetic Surgery; Depression; Diabetes; Dieting and Drugs for Dieting; Eating Disorders; Eczema; Genetic Disorders; Genetics, Genealogy and History; Low Carbohydrate Diets; Lupus; Fibromyalgia; Menopause; Mesothelioma and Lung Cancer; Muslims & Medical History; Osteoporosis; Pediatric Radiology; Perinatology; Periodic Paralysis; Pregnancy; Psoriasis; Rheumatoid Arthritis; Severe Acute Respiratory Syndrome (SARS Virus); Stress; Substance Abuse; Weight Loss Surgery; Zoonotic Diseases (Diseases passed from Animal to Human).*

Most of the link topics listed above could relate to either the NLM or NeLH, yet because many of them were NOT associated with the NeLH, there may have been an underlying regional or geographic component. For instance, we were not surprised that websites concerning *Cosmetic Surgery, Weight Loss Surgery, or Low Carbohydrate Diets* link to the NLM. Cosmetic surgery is advanced in the United States and has become an important aspect of American culture (e.g., Hollywood), as well as the low-carbohydrate eating plan (and all its health-related issues) made popular by the U.S. physician Dr. Robert C. Atkins. Also, not long ago, there was a serious SARS episode in North America (mainly in Toronto, Canada); therefore, it makes sense that one or two websites about *Severe Acute Respiratory Syndrome (SARS Virus)* link to the NLM. In 2003 George Bush gave health officials the authority to quarantine Americans infected with the highly contagious SARS, and many Americans began to think that this mystery illness made travel to Canada unsafe.

Aside from a regional component, it is obvious that some links were directed to the NLM because of its specialty resources (e.g., History of Medicine). The *Muslims & Medical History* topic came from an Islamic Medicine website (www.islamicmedicine.org/history.htm) with a directed link to the NLM's *Exhibitions and Public Programs* page concerning Islamic Culture and the Medical Arts.

A birds-eye view of our maps, Figures 3 and 4 (NeLH colink network) and Figures 7 and 8 (NLM colink network) reveals two very different configurations. The NLM is somewhat isolated from the rest of its colink network: more so than the NeLH is from its Web network. Also, the overall density of the NLM network is greater to a higher mean colink rate (see Tables IIA and IIIA in the Appendix). The NLM, unlike the NeLH, seems to be more frequently colinked with sites from the .edu domain than from any other domain, and it is also receiving more direct links from sites within its small Web network. We expect the NLM to fit within an educational environment due to its specialty resource on the History of Medicine and also due to its database for researchers interested in the Human Genome project.

A comparison between the second or top-level domain data shown in Figures 2 and 6 indicates that more US education sites (.edu) link to the NeLH than UK education sites (.ac.uk). This is an interesting finding since the NeLH has been developed to support the work and research habits of educated medical professionals throughout the UK. Where are the UK educational links and why are they missing from the NeLH data?

Last but not least, it is important to notice the international sites, which either link directly to the NLM or the NeLH or are situated within the same colink network. First, we can see that the World Health Organisation (who.int) appears on both the NeLH and NLM colink maps (see Figures 3, 4, 7, and 8). Also, the NLM is part of a Web network including direct links from sites in *Canada, Switzerland, Japan, Argentina, Norway and New Zealand*. By comparison, the NeLH has received inlinks from *Canada, Belgium, Australia, New Zealand and The Netherlands*. One might expect to see more European websites linking to the NeLH; however the presence of Commonwealth country sites like *Canada and Australia* make sense, particularly because their predominantly government funded, public healthcare systems are similar to the system in the U.K.

7. NeLH Service Delivery Implications

With the “Web intelligence” information that we have gathered using *LexiURL*, a few ideas have been generated concerning the NeLH’s service delivery. First, the NeLH is interested and will continue to be interested in satisfying the needs of its users; therefore link data are valuable when analyzed in depth because they give service managers an opportunity to learn a lot about the types of organisations and people that link to and possibly use their resource. For instance, the NeLH service manager who had originally viewed the colink maps was both surprised by and interested in the fact that many American Universities link to this digital library. Does this mean that an American user of the NeLH relies on this digital library for information not available in the U.S. or that there is a need to locate information complementary to what is available in the U.S.? Clearly we have some insight here leading to a future investigation concerning international users; a genuine “intelligence” finding.

Also, the mission of the NeLH is to reach out to and support the needs of healthcare practitioners across the U.K. Given that a training ground for these practitioners is the U.K. educational system perhaps there is a need to promote the digital library to users at the university level. In other words, an important service strategy might be to motivate a digital library use habit amongst U.K medical and nursing students, including other healthcare students in training, *before* they enter the National Health System, instead of hoping they will become regular users after.

At present, few studies provide substantial statistical evidence to show that numerous links to a particular part of a website is a strong indicator of frequent use (Brin & Page, 1998; Kleinberg, 1999). In this study, we found that the *Hitting the Headlines* resource at the NeLH website possessed a good number of direct links from other pages on the Web, but that there could in fact be more. Perhaps there are additional links, still hidden on the Web, but we believe that more links are worth creating. The value attached to this resource is that it appeals to both healthcare practitioners and patients alike; thus, it can lead to an interesting dialogue. With this resource, physicians are better able to respond to patients who approach them with questions about the medical information they have discovered through the news media, and patients can feel more confident that their physicians recognize the relationship between medical research and everyday practice. If further Web intelligence analyses could play a part in determining that numerous links to the *Hitting the Headlines* resource are an indicator of frequent use, then the NeLH would have confirmation of a specialized service that has become quite successful.

8. Conclusion

Is the Web intelligence analysis of a digital library worthwhile? According to this research, yes, it is. We have demonstrated that a *LexiURL* Web link analysis can provide interesting, if not surprising insights into who links to and potentially uses an important Web site. Nevertheless, if we consider the amount of time that is required to complete a full review of all the link/colink contexts, a researcher or digital library manager might only wish to carry out this analytic process 2 to 4 times per year. When

a log file analysis and a *LexiURL* link analysis are used together for comparative purposes, the findings are generally more complete: one can expect to find out where users are coming from (i.e., the Web referring domain), when they are using a site, and where potential users might discover a digital library site if they are currently not using it. Whilst a log file analysis of somebody else's Web site is normally not possible for a digital library webmaster, a link analysis can be conducted for any Web site; hence comparative link-based analyses are always possible. Finally, even though there are clear limitations inherent to using a search engine for data (i.e., incomplete coverage of the Web), the results returned are still useful as a clear indicator of how a network is growing within a significant portion of the Web.

9. Acknowledgements

The researchers would like to thank the director and digital library service manager, Dr. Ben Toth and Phil Vaughan of the National electronic Library for Health for providing the necessary background information and support for this case study. Funding for this research was granted in part by the Joint Information Systems Committee, United Kingdom.

10. References

- Adams, A. and Blandford, A. (2004), "The unseen and unacceptable face of digital libraries", *International Journal on Digital Libraries*, Vol. 4 No. 2, pp.71-81.
- Adams, A. and Blandford, A. (2005), "Digital libraries' support for the user's 'information journey', in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, Denver, Colorado, U.S.A., pp. 160-169.
- Atluri, V. and Ray, I. (2004), "Editors' message for the special issue on security", *International Journal on Digital Libraries*, Vol. 4 No. 3, pp. 137-138.
- Bar-Ilan, J. (2005), "What do we know about links and linking? A framework for studying links in academic environments", *Information Processing & Management*, Vol. 41, No. 3, pp.1973-1986.
- Beaulieu, A. (2005), "Sociable hyperlinks: an ethnographic approach to connectivity", in C. Hine (Ed.), *Virtual Methods: Issues in Social Research on the Internet*, Berg: New York, pp. 183-197.
- Björneborn, L., and Ingwersen, P. (2004), "Towards a basic framework for Webometrics", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 14, pp. 1216-1227.
- Blandford, A., Buchanan, G. and Jones, M. (2004), "Usability of digital libraries", *International Journal on Digital Libraries*, Vol. 4 No. 2, pp. 69-70.
- BMJ Clinical Evidence. How Clinical Evidence is put Together.* (2005), available at: http://www.clinicalevidence.com/ceweb/about/put_together.jsp.
- Borgman, C. L., and Furner, J. (2002), "Scholarly communication and bibliometrics", *Annual Review of Information Science and Technology*, Vol. 36, pp. 3-72.

- Borgman, C. L., Smart, L. J., Millwood, K. A., Finley, J. R., Champeny, L., Gilliland, A. J., and Leazer, G. H. (2005), "Comparing faculty information seeking in teaching and research: implications for the design of digital libraries", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 6, pp. 636-657.
- Brin, S., and Page, L. (1998), The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, Vol. 30 No. 1-7, pp. 107-117.
- Brody, T., Stamerjohanns, H., Harnad, S., Gingras, Y., Vallieres, F. and Oppenheim, C. (2004), "The effect of Open Access on Citation Impact", presented at: *National Policies on Open Access (OA) Provision for University Research Output: an International meeting*, Southampton University, Southampton UK, 19 February 2004, available at: <http://opcit.eprints.org/feb19oa/brody-impact.pdf>
- Chakrabarti, S. (2003), *Mining the Web: Analysis of Hypertext and Semi Structured data*, New York: Morgan Kaufmann.
- Cronin, B. (2005), "Vox populi: Civility in the blogosphere", *International Journal of Information Management*, Vol. 25 No. 6, pp. 549-550.
- Crow, R. (2002), "The case for institutional repositories: A SPARC position paper", available at: <http://www.arl.org/sparc/IR/ir.html>.
- Fox, E., and Urs, S. (2002), "Digital libraries", *Annual Review of Information Science and Technology*, Vol. 36, pp. 503-589.
- Gadd, E., Oppenheim, C. and Proberts, S. (2003a), "RoMEO Studies 1: The impact of copyright ownership on academic author self-archiving", *Journal of Documentation*, Vol. 53 No. 3, pp. 243-277.
- Gadd, E., Oppenheim, C. and Proberts, S. (2003b), "RoMEO Studies 2: How academics want to protect their open-access research papers", *Journal of Information Science*, Vol. 29 No. 5., pp. 333-356.
- Gadd, E., Oppenheim, C. and Proberts, S. (2003c), "RoMEO Studies 4: an analysis of journal publishers' copyright agreements", *Learned Publishing*, Vol. 16 No. 4, pp. 293-308.
- Harnad, S. and Brody, T. (2004), "Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals", *D-Lib Magazine*, Vol. 10 No. 6, available at: <http://www.dlib.org/dlib/june04/harnad/06harnad.html>.
- Heery, R. and Anderson, S. (2005), *Digital repositories review*, available at: http://www.jisc.ac.uk/uploaded_documents/rep-review-final-20050220.pdf.
- Jones, S., Cunningham, S. J., McNab, R. and Boddie. (2000), "A transaction log analysis of a digital library", *International Journal of Digital Librarianship*, Vol. 3 No. 1, pp. 152-169.
- Kaplan, N. R. and Nelson, M. L. (2000), "Determining the publication impact of a digital library", *Journal of the American Society for Information Science*, Vol. 51 No. 4, pp. 324-339.
- Kim, H. J. (2000), "Motivations for hyperlinking in scholarly electronic articles: a qualitative study", *Journal of the American Society for Information Science*, Vol 51 No. 10, pp. 887-899.
- Kleinberg, J. M. (1999), "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, Vol. 46 No. 5, pp. 604-632.

- Lau, K-N., Lee, K-H., Ho, Y., and Lam, P-Y. (2004), "Mining the web for business intelligence: homepage analysis in the internet era", *Database Marketing & Customer Strategy Mangement*, Vol. 12 No. 1, pp. 1741-2439.
- Levene, M. (2005), *An Introduction to Search Engines and Web Navigation*, London, Addison-Wesley.
- Marks, J. T., Hussell, J. A. T., McMahon, T. M. and Luce, R. E. (2005), "ActiveGraph: A digital library visualization tool", *International Journal on Digital Libraries*, Vol. 5 No. 1, pp. 57-69.
- Marshall, B., McDonald, D., Chen, H., and Chung, W. (2004), "EBizPort: Collecting and analyzing business intelligence information", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 10, pp. 873-891.
- McCain, K. (1990), "Mapping authors in intellectual space: a technical overview", *Journal of the American Society for Information Science*, Vol. 41 No. 6, pp. 433-443.
- McKay, D., Shukla, P., Hunt, R. and Cunningham, S. J. (2004), "Enhanced browsing in digital libraries: three new approaches to browsing in Greenstone", *International Journal on Digital Libraries*, 4, Vol. 4 No. 4, pp. 283-297.
- McLean, N. and Lynch, C. (2004), "Interoperability between library and information services and learning environments -- bridging the gaps. A joint white paper on behalf of the IMS Global Learning Consortium and the Coalition for Networked Information", available at: http://www.imsglobal.org/digitalrepositories/CNIandIMS_2004.pdf.
- National electronic Library for Health. (2005), available at: <http://www.nelh.nhs.uk/>.
- Nicholas, D., Huntington, P., Lievesley, N., and Wasti, A. (2000), "Evaluating consumer Website logs: a case study of The Times/The Sunday Times Website", *Journal of Information Science*, Vol. 51 No. 5, pp. 144-154.
- Nicholas, D., Huntington, P., Lievesley, N., and Withey, R. (1999), "Cracking the code: Web log analysis", *Online & CD-ROM Review*, Vol. 23 No. 5, pp. 263-269.
- Ortega-Priego, J. L., and Aguillo, I. (2005), "A Web map of the SCIC research centres: a comparative study of the cosine and Pearson's r", in P. Ingwersen and B. Larsen (Eds.), *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden: Karolinska University Press, pp. 197-204.
- Peters, T. A. (1993), "The history and development of transaction log analysis", *Library Hi Tech* 42, Vol. 11 No. 2, pp. 41-66.
- Ross, S. and Hedstrom, M. (2005), "Preservation research and sustainable digital libraries", *International Journal on Digital Libraries*, Vol. 5 No. 4, pp. 317-324.
- Small, H. (1973), "Co-citation in the scientific literature: A new measure of the relationship between two documents", *Journal of the American Society for Information Science*, Vol. 24 No. 4, pp. 265-269.
- Spink, A. and Jensen, B. J. (2004), *Web Search: Public Searching of the Web*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

- Swan, A. and Brown, S.N. (2004a), "Authors and open access publishing", *Learned Publishing*, Vol. 17 No. 3, pp. 219-224.
- Swan, A. and Brown, S.N. (2004b), *JISC/OSI Journal Authors Survey Report*, available at: http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf.
- Tibbo, H. R. (2003), "On the nature and importance of archiving in the digital age", *Advances in Computers*, Vol. 57, pp. 1-67.
- The Aerial Reconnaissance Archives*. (2004), available at: <http://www.evidenceincamera.co.uk/>.
- The National Archives*. (2005), available at: <http://www.nationalarchives.gov.uk/about/>.
- Thelwall, M. (2001), "Web Log File Analysis: Backlinks and Queries", *ASLIB Proceedings*, Vol. 53 Vol. 6, pp. 217-223.
- Thelwall, M. (2003), "What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation", *Information Research*, Vol. 8 No. 3, paper no. 151, available at: <http://informationr.net/ir/8-3/paper151.html>.
- Thelwall, M. (2004a), *Link Analysis: An Information Science Approach*. San Diego: Academic Press.
- Tang, R. and Thelwall, M. (2004b), "Patterns of national and international web inlinks to US academic departments: An analysis of disciplinary variations", *Scientometrics*, 60, Vol. 60 No. 3, pp. 475-485.
- Tuominen, K., Talja, S., and Savolainen, R. (2003), "Multiperspective digital libraries: The implications of constructionism for the development of digital libraries", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 6, pp. 561-569.
- UK Data Archive* (2005), available at: <http://www.data-archive.ac.uk/>.
- United States National Library of Medicine. About the National Library of Medicine's Web Site*. (2005), available at: <http://www.nlm.nih.gov/about/nlmweb.html>.
- Vaughan, L. (2004a), "Exploring website features for business information", *Scientometrics*, Vol. 61 No. 3, pp. 467-477.
- Vaughan, L. (2004b), "Web hyperlinks reflect business performance: A study of U.S. and Chinese IT companies", *Canadian Journal of Information and Library Science*, Vol. 28 No. 1, pp. 17-31.
- Vaughan, L. and You, J. (2005), "Mapping business competitive positions using Web co-link analysis", in P. Ingwersen and B. Larsen (Eds.), *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden: Karolinska University Press, pp. 534-542.
- Wasserman, S. and Faust, K (1994), *Social Network Analysis*, Cambridge, England: Cambridge University Press.
- Web Intelligence Newsletter. Issue 2*. (2003), available at: <http://wi-consortium.org/newsletter/issue2.html>.
- Wheeldon, R., and Levene, M. (2003), "The best trail algorithm for assisted navigation of Web sites" in *1st Latin American Web Congress (LA-WEB 2003)*, Sanitago, Chile, IEEE Computer Society, pp. 166-179.

- White, H. D., and Griffith, B. C. (1982), "Author co-citation: a literature measure of intellectual structure", *Journal of American Society for Information Science*, Vol. 32 No. 3, pp.163-172.
- Witten, I. H., Boddie, S. J., Bainbridge, D., and McNab, R. J. (2000), "Greenstone: A comprehensive open-source digital library software system", in *Proceedings of the Fifth ACM Conference on Digital libraries*, San Antonio, Texas, USA, ACM Press, pp. 113-121.
- Zhong, N., Liu, J., and Yao, Y. (Eds.). (2003). *Web Intelligence. Springer Information Systems Applications*, Tokyo, Japan, Springer.
- Zuccala, A. (2006, to appear). "Author Cocitation Analysis is to intellectual structure as Web Colink Analysis is to ...?", *Journal of the American Society for Information Science and Technology*.

APPENDIX

Table IA. *LexiURL* extraction list of inlinks (N=50) to the National electronic Library for Health.

<http://800ceoread.com/blog/archives/001025.html>
<http://answers.google.com/answers/threadview?id=254138>
<http://aoltvsearch.aol.com/cat.adp?id=95918>
<http://autumnseminar.blogdrive.com/>
http://blog.xrefer.com/2003_06_01_xrefer_archive.html
<http://bmj.bmjournals.com/uknews/news20020218.shtml>
<http://bmj.bmjournals.com/uknews/news20020830.shtml>
<http://bmj.bmjournals.com/uknews/news20021015.shtml>
<http://brcaprevention.evidencewatch.com/>
<http://calder.med.miami.edu/catalog/alpha/n.html>
http://calder.med.miami.edu/catalog/subject/evidence_based_medicine.html
<http://catless.ncl.ac.uk/bifurcated/rivets/41>
<http://child3.dietj.biz/index3.aspx>
<http://communitiesofpractice.with1click.com/>
<http://communitiespractice.with1click.com/>
<http://david.davies.name/weblog/picturegallery.html>
http://denham.typepad.com/km/2003/09/corporate_memor.html
http://denham.typepad.com/km/2005/01/selfguided_km_t.html
<http://en.wikipedia.org/wiki/amniocentesis>
http://en.wikipedia.org/wiki/down_syndrome
http://en.wikipedia.org/wiki/list_of_digital_library_projects
http://en.wikipedia.org/wiki/national_health_service
<http://eprintblog.crimsonblog.com/archives20030501.html>
<http://fmed2.uncu.edu.ar/index4-biblioteca.php?pagina=otrasbd>
<http://frassle.rura.org/directory/index?feed=277>
http://frassle.rura.org/external1690_socialnetworkanalysis2
<http://freelink.org/muthu>
<http://gerardedawe.homestead.com/100000welcomeswebpage~ns4.html>
<http://healthcareresources.blogspot.com/>
<http://healthwriting.com/articles.htm>
<http://heart.evidencewatch.com/>
http://highered.mcgraw-hill.com/sites/0072819472/student_view0/real_hot_e-commerce_projects.html
<http://home.cc.umanitoba.ca/~mtennen/bookmarks/bookmark.html>
<http://homepage.mac.com/srivastava/abhishek/page1.html>
<http://hormonal-infertility.eaa47.com/>
http://hyscience.typepad.com/hyscience/2005/03/udge_nixes_schi.html
<http://imavja.20m.com/pedscapes.htm>
http://kegliography.blogspot.com/2003_07_01_kegliography_archive.html
<http://kmblogs.com/public/item/85526>
<http://kmblogs.com/public/item/85528>
<http://kmblogs.com/public/item/85529>
<http://kml.uindy.edu/shirley/ot/databases.html>
<http://kmterms.persianblog.com/>
http://knowledgeaforethought.blogs.com/knowledge_aforethought/2004/09/managing_boids.html
<http://lacuna.us/links.php>
<http://lib.itg.be/biblinks.htm>
<http://liberatingtheliterature.blogspot.com/>
<http://lonestar9.pitas.com/>
<http://lupusnewslog.blogspot.com/>
<http://medicine.plosjournals.org/perlserv?request=get-document&doi=10.1371/journal.pmed.0010004>

Table IIA. Partial matrix (n=15) of raw colink counts between the www.nelh.nhs.uk web site and other sites (note: the full map includes 50 sites). Cells along the diagonal include the highest count from each column.

	nelh.nhs.uk	nih.gov	cdc.gov	ox.ac.uk	harvard.edu	ama-assn.org	washington.edu	who.int	geocities.com	stanford.edu	medscape.com	umich.edu	uiowa.edu	fda.gov	bbc.co.uk
nelh.nhs.uk	25	10	16	12	8	5	5	5	7	17	11	11	7	8	9
nih.gov	10	29	20	10	3	7	3	10	9	24	10	11	10	19	13
cdc.gov	16	20	49	15	14	12	7	16	20	33	21	16	15	16	19
ox.ac.uk	12	10	15	33	9	14	6	11	10	19	14	16	11	7	9
harvard.edu	8	3	14	9	34	11	13	11	11	13	17	11	6	2	8
ama-assn.org	5	7	12	14	11	27	11	10	5	12	12	10	12	4	7
washington.edu	5	3	7	6	13	11	26	5	5	7	10	3	6	2	7
who.int	5	10	16	11	11	10	5	33	12	18	11	10	5	6	12
geocities.com	7	9	20	10	11	5	5	12	35	15	9	10	5	7	9
stanford.edu	17	24	33	19	13	12	7	18	15	52	18	22	17	21	22
medscape.com	11	10	21	14	17	12	10	11	9	18	31	14	12	6	10
umich.edu	11	11	16	16	11	10	3	10	10	22	14	30	12	10	11
uiowa.edu	7	10	15	11	6	12	6	5	5	17	12	12	27	11	11
fda.gov	8	19	16	7	2	4	2	6	7	21	6	10	11	26	10
bbc.co.uk	9	13	19	9	8	7	7	12	9	22	10	11	11	10	27

Summary Statistics

$$50(50-1)/2 = 1225$$

Mean colink rate (over 50 different Web sites): 14

Range of raw colink counts: 2 (harvard.edu; washington.edu; fda.gov) and 110 (emedicine.com; ucla.edu)

Number of unique colink pairs not made: 0

Number of unique colink pairs made once: 0

Table IIIA. A partial matrix (n=15) of raw colink counts between the www.nlm.nih.gov web site and other sites (note: the full map includes 50 sites). Cells along the diagonal include the highest count from each column.

	nlm.nih.gov	cdc.gov	fda.gov	harvard.edu	umich.edu	stanford.edu	geocities.com	washington.edu	ama-assn.org	wisc.edu	loc.gov	cornell.edu	upenn.edu	berkeley.edu	uiowa.edu
nlm.nih.gov	41	22	24	12	23	18	27	11	13	21	11	12	14	21	29
cdc.gov	22	73	22	14	21	21	52	18	27	27	19	21	27	42	29
fda.gov	24	22	59	19	23	24	36	13	25	27	14	22	25	28	45
harvard.edu	12	14	19	42	13	32	19	9	25	26	11	22	22	16	18
umich.edu	23	21	23	13	51	23	30	28	17	21	10	14	15	20	29
stanford.edu	18	21	24	32	23	58	28	17	30	37	10	24	27	21	27
geocities.com	27	52	36	19	30	28	107	24	37	37	26	23	32	62	41
washington.edu	11	18	13	9	28	17	24	38	16	17	8	10	13	15	16
ama-assn.org	13	27	25	25	17	30	37	16	60	34	15	26	31	27	27
wisc.edu	21	27	27	26	21	37	37	17	34	64	20	27	31	30	29
loc.gov	11	19	14	11	10	10	26	8	15	20	41	10	13	22	16
cornell.edu	12	21	22	14	24	23	10	26	27	10	44	28	20	20	23
upenn.edu	14	27	25	22	15	27	32	13	31	31	13	28	53	26	27
berkeley.edu	21	42	28	16	20	21	62	15	27	30	22	20	26	92	31
uiowa.edu	29	29	45	18	29	27	41	16	27	29	16	23	27	31	76

Summary Statistics

$$50(50-1)/2 = 1225$$

Mean colink rate (over 50 different Web sites): 24

Range of raw colink counts: 2 (aol.com; jhu.edu) and 107 (geocities.com; mit.edu)

Number of unique colink pairs not made: 0

Number of unique colink pairs made once: 0