How Is My Web Community Developing? Monitoring Trends In Web Service Provision

Brian Kelly, UK Web Focus and Ian Peacock, WebWatch
UK Office for Library and Information Networking, University of Bath, Bath, BA2 7AY, UK

E-mail: b.kelly@ukoln.ac.uk i.peacock@ukoln.ac.uk
Phone: +44 1225 323943 +44 1225 323570
Fax: +44 1225 826838

Abstract

*As the World Wide Web continues its tremendous rate of development, providers of services on the Web have difficult decisions to make regarding the deployment of new technologies: should they begin deployment of technologies such as HTML 4.0, CSS 2, Java, Dublin Core metadata, etc., or should they wait until the technologies mature. This paper describes the use of a web auditing / profiling robot utility known as **WebWatch** which can help service providers by providing information on the uptake of technologies within particular communities. A description of use of the WebWatch software within the UK Higher Education community is given, together with a discussion of the findings.*

Introduction

The Beleaguered Webmaster

In the early days of the web life was easy for the webmaster, to use the popular, if politically-incorrect term. A simple text editor (typically vi or emacs for the Unix user or Notepad for the Windows users) or simple HTML authoring tool would suffice for creating web pages. Add a graphical tool for creating and editing images, and the webmaster could create a website which could make use of most of the web technologies which were widely deployed in around 1994.

These days, however, life is much more difficult. Competition between the browser software vendors has hastened the development of a wide range of web technologies, much of which, sadly, appears to suffer from interoperability problems. The web standards community, principally the World Wide Web Consortium, has developed a range of new or updated web protocols (see article by Brian Kelly elsewhere in this edition of the Journal of Documentation) although, again, there are reports of implementation problems.

As the web becomes increasingly used to support core business functions, rather than simply as a noticeboard managed by enthusiasts in the IT department, the webmaster faces pressures to begin deployment of new technologies. He, and the webmaster is often male, is often not in a position to say no and point out deployment and interoperability problems.

Web Monitoring Tools

Web auditing and monitoring tools can assist the beleaguered webmaster by providing information on the uptake of web technologies. Such tools can provide evidence on how widely deployed particular technologies are and how they are used. This information is, of course, of use to a number of communities such as policy-makers, funders, software developers, etc.

In this paper the authors describe the use of a web monitoring tool based on web robot software which can be used freely on the Web without any special authorisation. A description of the robot software which has been developed by the authors is given. The paper then reports on the use of the tools within one particular community – UK Higher Education – and interprets the results. The paper concludes by describing other ways in which web monitoring tools can be used.

Robot Software

Background

How big is the Web? Clearly in order to answer this question automated software must be used.

In 1993 the first attempt to answer the question was made. The World Wide Web Wanderer (WWWW) web robot was developed as an automated tool to automatically follow links on web pages in order to count the total number of resources to be found on the Web. In June 1993 the robot detected 130 web sites, which had grown to over 10,000 by December 1994 and 100,000 by January 1996 [1].

Since this initial survey was started, a number of other trawls have been carried out, although, due to the current size of the web, trawls of the entire Web tend nowadays to be carried out by large organisations which have the required disk and server capacity. The Open Text Corporation's trawl reported by Tim Bray at the WWW 5 conference [2] indicated that by November 1995 there were over 11 million unique URLs and over 223,000 unique web servers.

Robot Software

The *World Wide Web Wanderer* and the *Open Text Spider* are examples of web robots. A web robot can be regarded as an automated browser, which will sequentially retrieve

web resources. Unlike a browser, however, a robot is not designed to retrieve resources for viewing. Robots typically retrieve web resources for auditing purposes, as described above, for indexing or for checking (such as robot software to detect broken links).

The current generation of web crawlers is large. A glance at the Web Robots Pages [3] reveals a list of over 160 well-known robots. These robots are used for a variety of purposes including auditing and statistics (such as the Tcl W3 Robot [4] and the RBSE Spider [5]), indexing (the NWI Robot [6] and Harvest [7]), maintenance (Checkbot [8] and LinkWalker [9]) and mirroring (Templeton [10]).

Robot software can be regarded as automated web browsers. A potential problem with robot software is the danger of causing server or network overload by requesting too many resources in a short space of time. In order to overcome this problem the Robot Exclusion Protocol [11] has been developed. This is a method that allows web administrators to indicate to robots which parts of their site the robots should not visit.

The WebWatch Project

Background

The WebWatch project is funded by the BLRIC (the British Library Research and Innovation Centre). The project is based at UKOLN, University of Bath. The aims of the WebWatch project are:

- To develop robot software to gather information on usage of web technologies within a number of communities within the UK.
- To use the software to collect the data.
- To develop (if appropriate) and use analysis tools to provide statistical analyses of the data.
- To produce reports explaining the analyses.
- To make recommendations to appropriate bodies on the information collected.
- To publicise reports to relevant communities.

The WebWatch project began in August 1997.

WebWatch Robot Software

Following an initial survey of robot software it was decided to make use of the Harvest software. Harvest [12] is a software suite which is widely used within the worldwide research distributed indexing community. A slightly modified version of the software

was used in the initial WebWatch trawl carried out in October 1997 across UK public library websites [13].

Once the data for this community and a number of other small trawls had been collected and analysed it became apparent that Harvest was very limited as an auditing robot. As it had been designed for indexing web resources, it did not allow non-textual resources, such as images, to be downloaded. Also as it processed the file suffix for web resources, rather than Internet MIME types, it was not possible to analyse resources by MIME types. In the light of these limitations and the difficulties found in extending Harvest it was designed to write our own WebWatch robot which would be designed for auditing purposes.

The current version of the WebWatch robot is written in perl5 and builds on previous versions.

Survey of UK Higher Education Entry Pages

In October 1997 a WebWatch trawl of UK University entry pages was carried out. The trawl was repeated on 31 July 1998 (which terminated on 2 August). The initial results have been published elsewhere [14]. In this paper we give a brief summary of the original survey, a more detailed report of the second trawl and a comparison between the two trawls.

Initial Trawl of UK Universities

The initial trawl of UK University entry pages began on the evening of Friday 24th October 1997. The WebWatch robot analysed the institutional web entry point for UK Universities and Colleges as defined in the HESA list [15]. This list contained the entry points for 164 institutions. The WebWatch robot successfully trawled 158 institutions. Six institutional home pages could not be accessed, due to server problems, network problems or errors in the input data file.

Second Trawl of UK Universities

The second trawl of UK University entry points was initiated on the evening of Friday 31 July 1998. This time the NISS list of Higher Education Universities and Colleges [17] was used for initial trawl. This file contains 170 institutions. The WebWatch robot successfully trawled 149 institutions. Twenty-one institutional home pages could not be accessed, due to server problems, network problems, restrictions imposed by the robot exclusion protocols or errors in the input data file.

A total of 59 sites had `robots.txt` files. Of these, two sites (Edinburgh and Liverpool universities) prohibited access to most robots. As these sites were not trawled they are excluded from most of the summaries. However details about the server configuration is included in the summaries.

Note that when manually analysing outliers in the data it was sometimes found that information could be obtained which was not available in the data collected by the robot.

A brief summary of the findings is given below. More detailed commentary is given later in this article.

Server	**Usage (No. / %) Oct 1997**	**Usage (No. / %) July 98**	**Comments**
Apache	48 / 31%	62 / 42%	Mostly Unix platform (possibly also Windows NT)
Netscape	24 / 15%	25 / 17%	Unix and Windows NT platforms
Microsoft	13 / 8%	20 / 13%	Windows NT platform
NCSA	33 21%	14 / 9%	Unix platform
CERN	20 / 13%	13 / 9%	Unix platform
Webstar	3 / 2%	4 / 2%	Macintosh platform
Novell	3 / 2%	3 / 2%	PC
OSU	5 / 3%	2 / 1%	Dec VMS platform. Used at http://www.mdx.ac.uk/ and http://www.rhbnc.ac.uk/
Lotus Domino	0 / 0%	1 / 1%	Windows NT platform. Used at http://www.henleymc.ac.uk/
BorderWare	2 / 1%	1 / 1%	Used at http://www.marjon.ac.uk/
SWS	0 / 0%	1 / 1%	Sun (Unix) platform. Used at http://www.norcol.ac.uk/
HTTPS	1 / 1%	1 / 1%	Used at http://www.rgu.ac.uk/
WinHTTPD	1 / 1%	1 / 1%	Used at http://www.ssees.ac.uk/
WN	0 / 0%	1 / 1%	Used at http://www.haac.ac.uk/
Microsoft PWS	1 / 1%	0 / 0%	Was used at http://www.rave.ac.uk/. Now upgraded to Microsoft-IIS.
Purveyor	1 / 1%	0 / 0%	Was used at http://www.uwic.ac.uk/. Now upgraded to Microsoft-IIS
Roxen Challenger	1 / 1%	0 / 0%	Used at http://www.uel.ac.uk/. Server down at time of second trawl.
WebSite	1 / 1%	0 / 0%	Used at http://www.york.biosis.org/. Site not in input file of second trawl.
TOTAL	**157**	**149**	

Table 1 Table of Server Usage

As can be seen from Table 1 the Apache server has grown in popularity. This has been mainly at the expense of the NCSA and CERN servers, which are now very dated and no longer being developed. In addition a number of servers appear to be no longer in use within the community (e.g. Purveyor and WebSite). Microsoft's server has also grown in popularity.

The popularity of Apache is also shown in the August 1998 Netcraft Web Server Survey [16], which finds Apache to be the most widely used server followed by Microsoft-IIS and Netscape-Enterprise. The Netcraft surveys are taken over a wider community than the academic sites looked at in this paper. The community surveyed by Netcraft is likely to consist of more diverse platforms (such as PCs) whereas academic sites show a bias towards Unix systems. This may explain the differences in the results of the next most popular servers.

Table 2 shows a profile of HTTP headers.

HTTP/1.0	50%
HTTP/1.1	50%
Cachable resources	54% of HTML pages and 60% of images
Non-cachable resources	1% of HTML pages and 0% of images
Cachability not determined	36% of HTML pages and 40% of images

Table 2 HTTP Headers

Note that this information was not collected for the first trawl due to limitations in the robot software.

In Table 2 a resource is defined as cachable if:

- It contains an `Expires` header showing that the resource has not expired
- It contains a `Last-Modified` header with a modification date greater than 1 day prior to the robot trawl.
- It contains the `Cache-control: public` header

A resource is defined as **not** cachable if:

- It contains an `Expires` header showing that the resource has expired
- It contains a `Last-Modified` header with a modification date coinciding with the day of the robot trawl

- It contains the `Cache-control: no-cache` or `Cache-control: no-store` headers
- It contains the `Pragma: nocache` header

The cachability of resources was not determined if the resource used the `Etag` HTTP/1.1 header, since this would require additional testing at the time of the trawl which was not carried out.

Figure 1 gives a histogram of the total size of the institutional entry point.

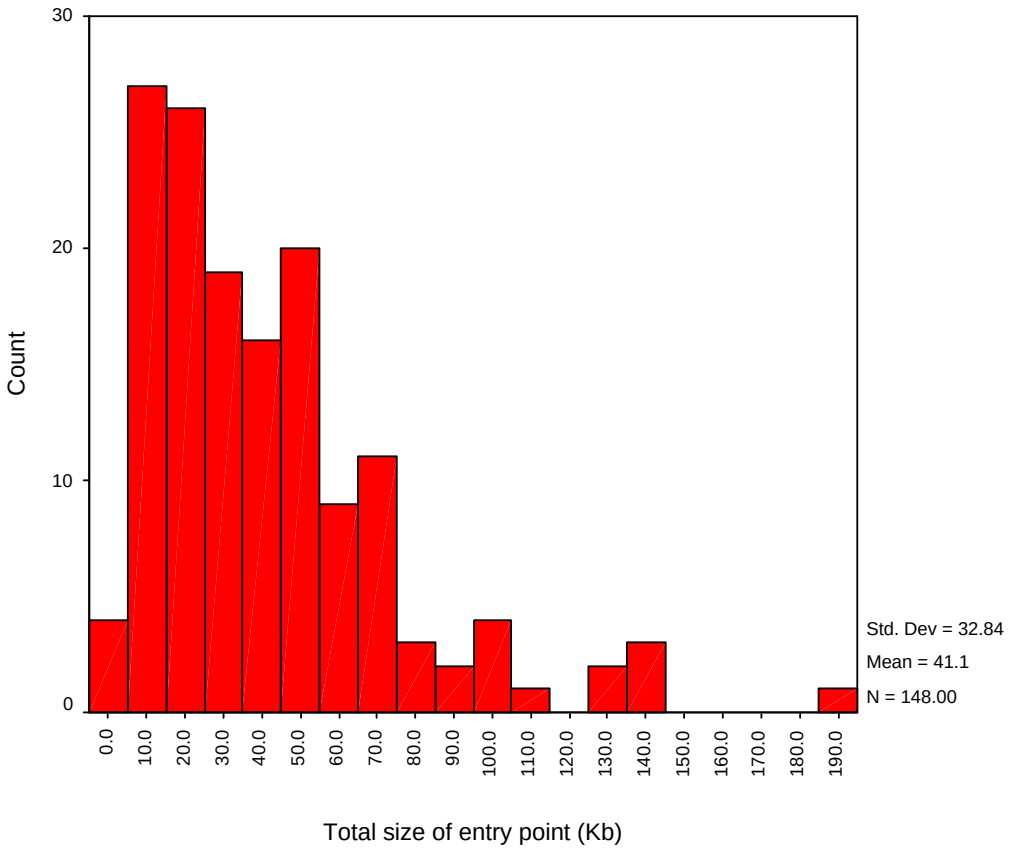

Figure 1 Size of Entry Point

As shown in Figure 1, four institutions appear to have an institutional web page which is less than 5Kbytes. The mean size is 41 Kb, with a mode of 10-20 Kb. The largest entry point is 193 Kbyes.

Note that this information is based on the size of the HTML file, any framed or refresh HTML pages, inline images and embedded Java applets.

It does not include any background images, since the current version of the robot does not parse the <BODY> element for the BACKGROUND attribute. Subsequent analysis showed that 56 institutions used the BACKGROUND attribute in the <BODY> element. Although this would increase the file size, it is unlikely to do so significantly as background elements are typically small files.

The histogram also does not include any linked style sheet files. The WebWatch robot does not parse the HTML document for linked style sheets. In this the robot can be regarded as emulating a Netscape 3 browser.

Figure 2 gives a histogram for the number of images on the institutional entry point. As mentioned previously this does not include any background images.

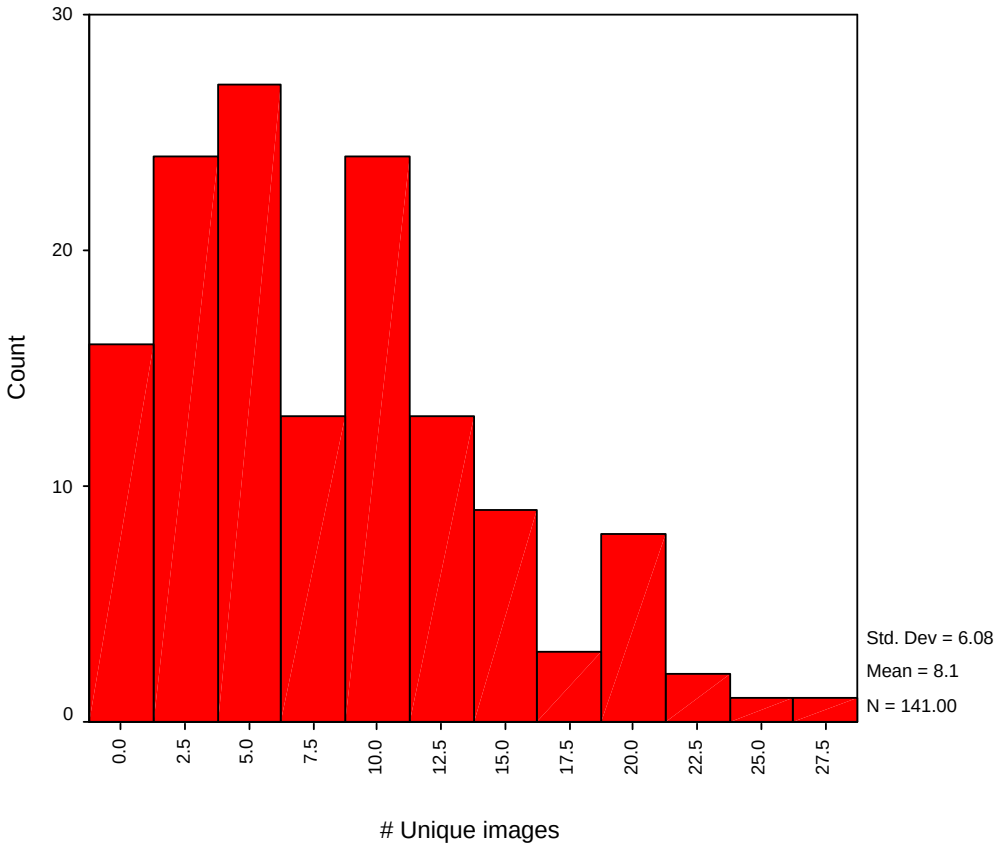

Figure 2 Numbers of Images

Figure 3 gives a histogram for the number of hypertext links from institutional entry points.

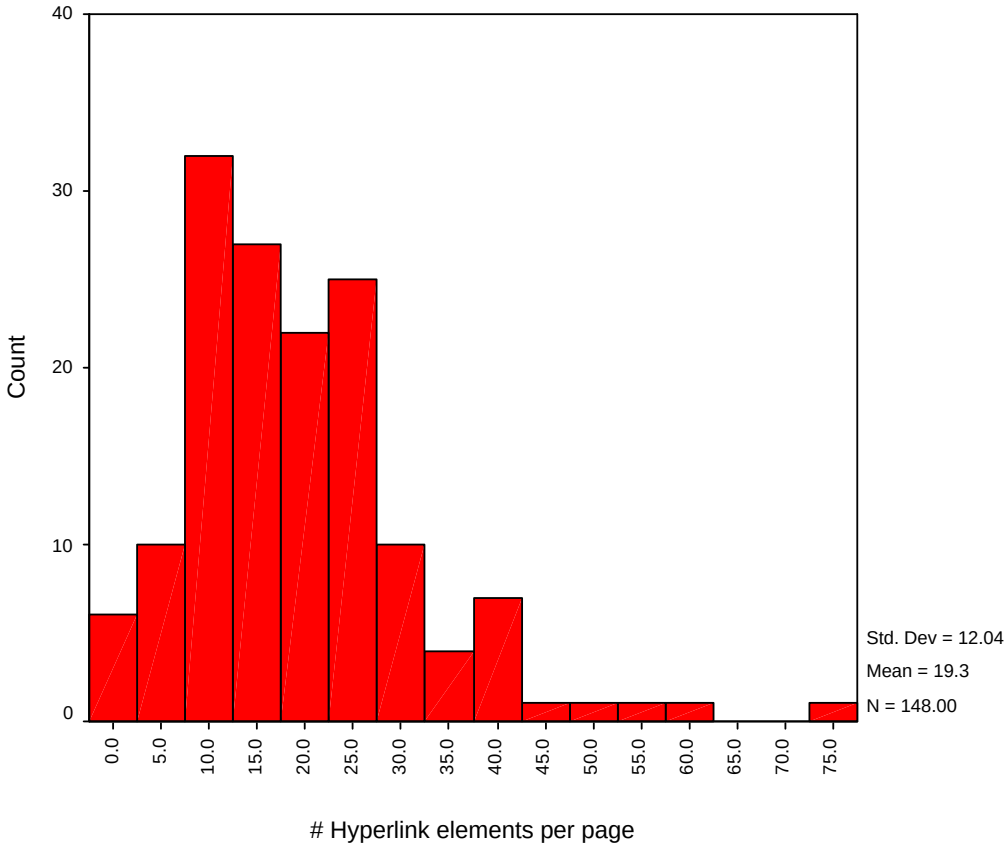

Figure 3 Link Profiles.

Note that Figure 3 gives the total number of links which were found. This includes <A> elements and client-side image maps. Note that typically links in client-side maps are duplicated using the <A> element. No attempt has been made in this report to count the number of unique links.

Discussion of Findings

In this section we discuss the findings of the trawls.

The discussion covers the accessibility of the pages and the technologies used. In the accessibility discussion we consider factors relevant to users accessing the pages, including the files sizes (which affects download times), whether the pages can be cached (which also affects download times) and the usage of hyperlinks (which can affect the usability). In the technology discussion we consider the technologies used, such as server hardware and software, and web technologies such as use of JavaScript and Java, metadata and stylesheets.

The results of the WebWatch trawl are intended to correspond closely with those that would be observed by a user using a web browsers. This is unlike, for example, many indexing robots which are not capable of processing frames. Robot software can also have problems in downloading linked resources, such as style sheet files, parsing HTML elements which may link to external resources, such as images, or processing HTTP headers, such as redirects. Robots developers often have a conservative approach to implementing new features in order to minimise the dangers of robots recursively requesting resources or causing other network or server problems.

The WebWatch has a similar conservative design. In a number of cases the automated analyses were modified by subsequent manual investigation in order to provide results which are applicable to a human view of a website (for example the size of a framed resource is the sum of the framed elements and not the hosting frameset). Where it has not been possible to do this, commentary is provided.

Size of Institutional Entry Point

The majority of institutional entry points appear to be between 10 Kb and 100 Kb (excluding background images which, as stated previously, were not included in the analysis).

Details of the largest and smallest institutional entry points are given in Table 3.

Institution	Size	Comments
South Devon College http://www.torbay.gov.uk/sdc/	0.5 Kb	Error in input data file. Points to directory listing, not to resource
Royal College of Music http://www.rcm.ac.uk/	2.9 Kb	
Westminster College http://www.ox-west.ac.uk/	3.9 Kb	Temporary interface while website being redesigned
University of Plymouth http://www.plym.ac.uk/	4.2 Kb	Contains background image (size not included in analysis)
Kent Institute of Art and Design http://www.kiad.ac.uk/	192 Kb	Contains animated GIF
University of Greenwich http://www.gre.ac.uk/	145 Kb	Contains animated GIF
Regent's College http://www.regents.ac.uk/	143 Kb	(Not available for manual analysis at time of writing)

University of Central England http://www.uce.ac.uk/	137 Kb	Contains animated GIF
King Alfred's http://www.wkac.ac.uk/	134 Kb	Contains animated GIF

Table 3 Summary Details of Largest and Smallest Sites in Current Trawl

Although perhaps not noticeable when accessing the page locally or across the SuperJANET network the large differences in sizes between, for example, the entry points for the University of Plymouth University and the Kent Institute of Art and Design are likely to cause noticeable differences in the download time for overseas users or accesses using modems.

It was also noted that all of the large sites which were available for manual inspection contained animated images.

Cachability of Institutional Entry Point

Interest in caching has grown in the UK Higher Education community since the advent of institutional changing for international bandwidth. In addition to interest in the cachability of resources from overseas websites, institutions are interest in the cachability of their own pages, especially key pages such the main entry point. Speedy access to such pages from local caches can be important when attempting to provide information to remote users, such as potential students. Unfortunately the need to provide cache-friendly pages may conflict with the need to provide attractive customised pages.

A study of the cachability of institutional entry points was carried out in order to observe the priorities given by institutions.

Over half of the institutional entry points have been found to be cachable, and only 1% not-cachable. 40% of the HTML resources used the `Etag` HTTP/1.1 header which is the current recommended method of establishing cachability. Unfortunately in order to identify if a resource can be cached the `Etag` value needs to be rechecked on a subsequent trawl and this was not carried out during this survey.

Links from Institutional Entry Point

The histogram of the numbers of hyperlinks from institutional entry points shows an approximately normal distribution, with a number of outliers indicating a small number of institutions with a large number of links. The institutional with the largest number of links on its entry point was Royal Holloway at <URL: `http://www.rhbnc.ac.uk/`>. The entry point contained 76 hyperlinks.

Providing a simple, uncluttered interface, especially to users accessing an institutional entry point for the first time, is arguably preferable to providing a comprehensive set of links to resources, although it could be argued that the a comprehensive set of links can minimise the navigation though series of sub-menus.

Future WebWatch trawls of institutional entry points will monitor the profile of hyperlink usage in order to determine any interesting trends.

"Splash Screens"

"Splash screens" are pages which are displayed for a short period before an alternative page is displayed. In the commercial world splash screens are used to typically used to display some form of advertisement before the main entry page, containing access to the main website , is displayed. Splash screens are normally implemented using the `<META REFRESH="value">` element. Typically values of about 5 seconds are used. After this period the second page is displayed.

In the initial WebWatch trawl, a total of five occurrences of the `<META REFRESH="value">` element were found. Of these, two had a value of 0. This provides a "redirect" to another page rather than displaying a splash screen.

In the second WebWatch trawl, a total of four occurrences were found (at the universities of Glamorgan, Greenwich, Sheffield and Staffordshire). Further investigation revealed that a number of additional sites use this feature which weren't detected in the robot trawl, due to the site being unavailable at the time of the trawl.

Further details are given in Table 4.

Institution	**Trawl Oct 97**	**Trawl July 98**
De Montford University	Refreshes after 8 seconds	Refreshes after 8 seconds
Glasgow School of Art	Redirects after 10 seconds	Redirects after 10 seconds (Note site not trawled due to omission in input file)
Glamorgan	Redirects to static page	Redirects to static page
Greenwich	Redirect to static page containing server-side include	Redirect to static page containing server-side include
Queen's University Belfast	Refreshes after 10 minutes	No refresh
Ravensbourne College of Art and Design	No refresh	Redirect (Note site not trawled due to omission in input file)

Sheffield	No refresh	Refresh after 10 minutes
Staffordshire	No refresh	Redirect to CGI script

Table 4 Comparison of Client-Side Refreshes

Metadata

Metadata can aid the accessibility of a web resource by making the resource more easy to find. Although the management of metadata may be difficult for large websites, management of metadata for a single, key page such as the institutional entry point should not provide significant maintenance problems.

The main HTML elements which have been widely used for resource discovery metadata are the <META NAME="keywords" VALUE="…"> and <META NAME="description" VALUE="…">. These elements are supported by popular search engines such as Alta Vista.

The resource discovery community has invested much time and energy into the development of the Dublin Core attributes for resource discovery. However as yet no major search engine is making use of Dublin Core metadata.

Metadata Type	**Oct 1997**	**Jul 1998**
Alta Vista metadata	54	74
Dublin Core	2	2

Table 5 Use of Metadata

As can be seen from Table 5, the metadata popularised by Alta Vista is widely used, although perhaps not as widely used as might have been expected, given the ease of creating this information on a single page and the importance it has in ensuring the page can be found using the most widely used search engines.

Dublin Core metadata, however, is only used on two institutional entry points: the University of Napier and St George's Hospital Medical School. Although this may be felt to be surprising given the widespread awareness of Dublin Core within the UK Higher Education community, the very limited use appears to be indicative that web technologies are not used unless applications are available which make use of the technologies.

Server Profiles

Since the initial trawl the server profile has changed somewhat. A number of server which were in use in October 1997 (Purveyor, BorderWare, WebSite, Roxen Challenger, Windows PWS) have disappeared. The major growth has been in usage of Apache, which has grown in usage from 31% to 42%.

Unfortunately it is not possible to obtain the hardware platform on which the server is running. Certain assumptions can be made. For example, Apache probably runs on Unix platforms since the Windows NT version is relatively new and reports indicate that the Windows NT version is not particularly fast. The Microsoft IIS server probably runs on a Windows NT platform. The CERN and NCSA server probably run on Unix. Unfortunately it is difficult to make realistic assumptions about the Netscape servers since these have been available for Unix and Windows NT platforms for some time.

Based on these assumptions Table 6 gives estimates for platform usage, based on the Netscape server being used solely on Unix or Windows NT.

Platform	**Estimated Min.**	**Estimated Max.**
Unix	89	115
Windows NT	21	46
Other PC platform	6	6
Macintosh	4	4
DEC	2	2

Table 6 Estimated Platform Usage

As may be expected the Unix platform is almost certainly the most popular platform. (This cannot be guaranteed, since the Apache server is now available for Windows NT. However as it has only been available on Windows NT for a short period and the Windows NT version is believed to be less powerful than Microsoft's IIS server, which is bundled free with Windows NT, it appears unlikely that Apache has made much inroads in the Windows NT world).

It will be interesting to analyse these results in a year's time, to see, for example, if Windows NT gains in popularity.

Java

None of the sites which were trawled contained any <APPLET>, <OBJECT> or <EMBED> elements, which are used to define Java applets. However it had been previously noted that the Liverpool University entry point contained a Java applet. Inspection of the `robots.txt` file for this site showed that all robots except the Harvest robot were excluded from this site.

The little use of Java could be indicative that Java does not have a role to play in institutional entry points or that institutions do not feel that sufficient number of their end users have browsers which support Java. The latter argument does, however, appear to contradict the growing use of technologies such as Frames and JavaScript which do require modern browsers.

JavaScript

In the initial trawl 22 of the 158 sites (14%) contained a client-side scripting language, such as JavaScript. In the second trawl 38 of the 149 sites (26%) contained a client-side scripting language, such as JavaScript.

The increasing uptake would appear to indicate confidence in the use of JavaScript as a mainstream language and that incompatibility problems between different browsers, or different versions of the same browser are no longer of concern.

With the increasing importance of client-side scripting languages in providing responsive navigational aids we can expect to see even more usage in the future. Future WebWatch trawls will help to identify if this supposition is true.

Frames

There has been a small increase in the number of sites using frames. In the original trawl 12 sites (10%) used frames. In the second trawl a total of 19 (12%) sites used frames.

HTML Validation

In the second trawl only three sites contained a page of HTML that validated without errors against the HTML3.2 DTD. Since it is reasonable to assume that most institutional webmasters are aware of the importance of HTML validity and have ready access to HTML validators (such as the HTML validation service which is mirrored at HENSA [17]) we might recommend a greater adoption of validated HTML pages.

Future Work

The WebWatch project has developed and used robot software for auditing particular web communities. Future work which the authors would like to carry out include:

- Running regular trawls across consistent samples in order to provide better evidence of trends.

- Making the data accessible for analysis by others across the Web. This would probably involve the development of a backend database which is integrated with the Web, enabling both standard and *ad hoc* queries to be initiated.

- Developing a number of standardised analyses. For example the development of an analysis system for analysing the accessibility of a website for the visually impaired, or the cachability of a website.

- Providing a web-based front-end for initiating “mini-WebWatch” analyses. Work on this has already begun, with the release of a web form for analysing HTTP headers [18].

References

[1] Gray, M. 1995. *Measuring the Growth of the Web*. See http://www.mit.edu:8001/people/mkgray/growth/ (visited 3 August 1998).

[2] Bray, T. 1996. *Measuring The Web*. Computer Networks and ISDN Systems Vol. 28, Nos. 7-11.

[3] The Web Robots Pages. http://info.webcrawler.com/mak/projects/robots/robots.html (visited 3 August 1998).

[4] Tcl W3 Robot. http://hplyot.obspm.fr/~dl/robo.html (visited 3 August 1998).

[5] RBSE Spider. http://rbse.jsc.nasa.gov/eichmann/urlsearch.html (visited 3 August 1998).

[6] NWI Robot. http://www.ub2.lu.se/NNC/projects/NWI/the_nwi_robot.html (visited 3 August 1998).

[7] Harvest. http://harvest.cs.colorado.edu/ (visited 3 August 1998).

[8] Checkbot. http://www.xs4all.nl/~graaff/checkbot/ (visited 3 August 1998).

[9] LinkWalker. http://www.seventwentyfour.com/ (visited 3 August 1998).

[10] Templeton. http://www.bmtmicro.com/catalog/tton/ (visited 3 August 1998).

[11] Robots Exclusion. http://info.webcrawler.com/mak/projects/robots/exclusion.html (visited 3 August 1998).

[12] Harvest Web Indexing. http://www.tardis.ed.ac.uk/harvest/ (visited 3 August 1998).

[13] Kelly, B., Ormes, S.L. and Peacock, I. *Robot Seeks Public Library Websites*. LA Record Dec 1997 Vol 12 (99).

[14] Kelly, B. 1997 *WebWatching UK Universities and Colleges*. http://www.ariadne.ac.uk/issue12/web-focus/ (visited 3 August 1998).

[15] HESA, *HESA List of Higher Education Universities and Colleges*. http://www.hesa.ac.uk/hesect/he_inst.htm (visited 3 August 1998).

[16] Netcraft, http://www.netcraft.co.uk/ (visited 20 August 1998).

[17] NISS, *Higher Education Universities and Colleges*. http://www.niss.ac.uk/education/hesites/cwis.html (visited 3 August 1998).

[18] WebTechs, *HTML Validation Service*. http://www.hensa.ac.uk/html-val-svc/ (visited 7 August 1998).

[19] UKOLN, *URL-info*. http://www.ukoln.ac.uk/web-focus/webwatch/services/url-info/ (visited 3 August 1998).