A Fair History of the Web? Examining Country Balance in the Internet Archive[1]

Mike Thelwall
School of Computing and Information Technology, University of Wolverhampton,
35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK
E-mail: m.thelwall@wlv.ac.uk

Liwen Vaughan
Faculty of Information and Media Studies
University of Western Ontario
London, Ontario, N6A 5B7, Canada
E-mail: lvaughan@uwo.ca

Abstract

The Internet Archive, an important initiative that maintains a record of the evolving Web, has the promise of being a key resource for historians and those who study the Web itself. The Archive's goal is to index the whole Web without making any judgments about which pages are worth saving. The potential importance of the Archive for longitudinal and historical Web research leads to the need to evaluate its coverage. This article focuses upon whether there is an international bias in its coverage. The results show that there are indeed large national differences in the Archive's coverage of the Web. A subsequent statistical analysis found differing national average site ages and hyperlink structures to be plausible explanations for this uneven coverage. Although the bias is unintentional, researchers using the Archive in the future need to be aware of this problem.

Introduction

The Internet Archive (www.archive.org) is a unique digital library that combines modern technology with the ancient practice of archiving to provide an international publicly-accessible resource of immense value. The Archive's objective is to store in perpetuity huge collections of digital information, an important and highly significant mission (Chavez-Demoulin, Roehrl, Roehrl, & Weinberg, 2000; Council on Library and Information Resources, 2002; Featherstone, 2000). Although it is not the only archive on the Web (Masanès, 2002; Rauber, Bruckner, Aschenbrenner, Witvoet & Kaiser, 2002), it is the biggest. The Archive is particularly important for scholarly communication, because of the increasing use of Web resources in teaching and academic research (Kenney, McGovern, Botticelli, Entlich, Lagoze & Payette, 2002).

The Archive is the invention of one man, Brewster Kahle, and has been funded predominantly from U.S. sources, but it remains open and freely available to anyone with Web access (Internet Archive, 2002a; Koman, 2002). In common with all digital libraries, and in fact with all libraries, the Archive acquires, organizes and disseminates information using contemporary technologies (Fox & Urs, 2002). Physically, the Archive is a large collection of computers in a building in San Francisco, staffed by highly skilled programmers. Although it holds a variety of resources, including digitized old films and a large text archive, its flagship project is a historical archive of the Web since 1996. This is an enormous collection of Web pages that have been obtained from the Alexa Web crawler (www.alexa.com). The particular differentiating feature for this mission is that it keeps all retrieved copies of Web pages (indexed by their URL) so that changes in a page over time can be tracked, and old pages that have been deleted from the Web can still be found. The resource has been found useful for several academic research projects (Hawking, Craswell & Thistlewaite, 1999; Suel, & Yuan, 2001; Vaughan & Thelwall, 2003a). The Web archive is a modern project in scope: it is not an attempt to

[1] Library and Information Science Research, to appear in 2004.

identify and save only the important Web pages, but to save *all* of them. The archiving of objects that are not recognized as being of cultural importance during their normal lifetime is now seen as an important activity in a way that it has not previously been (Eagleton, 1996, chapter 1; Lyman, 2002; Osborne, 1999). Other Web archives follow a more selective site indexing policy (Masanès, 2002).

As the only potential source for retrospective coverage of a large area of the Web, at least three important types of scholarly activities may need to rely upon it in the future. First, online resources are increasingly cited in academic papers, and the Archive is a logical first place to look when a cited resource has disappeared. Second, longitudinal studies of the Web can be conducted retrospectively via the Archive. According to Rousseau (1999), "Collecting time series should be an essential part of Internet research". The Archive avoids the need for a delay between formulating a question about change in the Web over time and the collection of sufficient data to investigate it. Currently, this kind of research takes months or years to collect the data (Koehler, 1999; McMillan, 2001; Rousseau, 1999) or uses commercial search engine date parameters (Leydesdorff & Curran, 2000; Leydesdorff, 2001; Uberti, 2003), which only yield pages that have not changed since the period specified. Third, the Web is of immense historical significance and historians will wish to study periods in its growth.

At the time of writing (the beginning of 2003), there were two methods to access archived pages. A Web-based search interface (the Wayback Machine) provides easy access, (www.archive.org). With this facility, users enter an URL and the Archive returns a table of information detailing all its copies of the page, along with archiving dates. The list is clickable to retrieve the individual pages. No facility is provided to search by page text, as would be standard for commercial search engines. Ideally, it would be useful to be able to submit queries looking for the first pages to include any given text, but the software to operate such queries would be highly complex. A second route to access the Archive is available to those with programming skills, which is to gain permission to use the archive's computers to access the data files directly. Such permission is freely given to researchers.

The Archive is clearly a resource of long-term international importance, the Web's only universal archival library recording the history of the Web. Both politicians and historians recognize the significance of archives in the chronicling and interpreting the past. For example, politicians sometimes destroy archives as part of a process of removing the cultural background of a people, and the details of archiving and particular archives are the subject of repeated discussions in the humanities (Brown & Davis-Brown, 1998; Derrida, 1996; Featherstone, 2000). Although much has been written about how the way in which archives are designed alters perceptions of the past (Ernst, 1999; Osborne, 1999), the concern here is not with the *nature* of the impact that the Internet Archive could have, only its coverage and bias. The importance of the Archive means that it is necessary to evaluate its claims and services stringently and critically. The fundamental question is, 'Which materials are preserved in the archive and which are excluded?" (Brown & Davis-Brown, 1998).

Problem Statement

Techniques for evaluating digital libraries and archives differ from those for traditional ones in terms of methodology, but the common objective is to assess the ability of the Archive to provide pertinent information to users. The evaluation will not attempt to be comprehensive but focuses upon the international coverage of the Archive and whether there is unbiased reporting of different countries. If researchers use the Archive as an information source, they need to know as much as possible about its coverage, given its aim to archive the whole Web. Search engines do not (Lawrence & Giles, 1999) and cannot (Thelwall, 2002) index the whole Web, so the Archive will not index the whole Web. Is the Archive's coverage of the Web biased, and if so, in what way? The rationale behind probing this question is that a fair and balanced history of the Web requires a fair representation of Web sites of different countries and that this is a critical question for any researcher seeking to use the Archive for data that crosses national boundaries. Given the international nature of the Web the choice of this issue

for bias evaluation is natural but evidence of significant bias would also be of interest to nation-based studies too: presumably, if there is one kind of bias then there will be others. Two specific research questions are investigated in this study.

- Are there national biases in the coverage of the Internet archive? More specifically, does it cover a different percentage of the Web sites from different countries?
- If national biases are found, can they be explained by technical factors inherent in the Web? In particular, is the distribution of Web links a key factor?

Literature Review

The Internet Archive

Numerous online articles about the Archive are available, many based upon interviews with its founder (e.g. Green, 2002; Internet Archive, 2002b; Koman, 2002; Wilson, 2001). These have tended to focus on the technical requirements for the archive, its implementation in practice and its founder's vision of the future. An overview of the crawling process (apparently before the crawling task was separated off as Alexa) is discussed elsewhere (Burner, 1997), but the details of the operation of the Archive's crawler are not of significance for this study. Definitive confirmation of the use of links by the crawler to find new sites is given, however. "The list of sites would be built by collecting "external" references from Web pages (those references pointing off-site)" (Burner, 1997). Legal issues concerning problems in archiving copyright material on the Web are important for the Archive (Lyman, 2002), but they are outside the scope of this project.

Search Engines

Since a search engine provides the pages for the Archive, it is pertinent to examine the literature describing search engine coverage issues. The Lawrence and Giles (1999) study of actual coverage by search engines has shown maximum coverage to be an estimated 16% of the publicly indexable pages – those do not have authentication requirements or protected by "robots exclusion standard".

There are two fundamental problems with crawling the Web that can explain partial coverage by crawlers. First, pages cannot be found if their URLs are not known, so pages that are not linked to are likely to be effectively invisible. Second, Web servers can create pages in response to requests, which means that the Web is theoretically virtually infinite in size. As a result of this, search engine crawlers must either have human intervention or a heuristic to stop crawling sites that appear to be dynamically serving pages without limits. A concrete example of this has been reported for a New Zealand university Web site (Thelwall & Wilkinson, 2003).

Some previous studies have analyzed the international coverage of commercial search engines. Major differences were found in 1999 for a range of 42 countries with Yahoo!, HotBot, AltaVista, MSN, and InfoSeek. Technical differences in HTML design were cited as a possible explanation (Thelwall, 2000). A recent study attempted a more detailed comparison for a smaller range of four countries, finding evidence that national differences in counts of links to a site indexed by a search engine would explain the differentiated coverage (Vaughan & Thelwall, 2003b). Differences in links indexed could be either a genuine national difference or a historical artifact of a later use of the Web by countries such as China.

The issue of technical obstacles for the indexing of Web pages (Thelwall, 2002) requires more explanation. There are two related issues: the ability of search engines to find pages through links and their ability to index them. The first is the main concern here. Although there are many pages that are not linked to by a link path from the site home page, there are other pages that are connected in this way but, nevertheless, will not be indexed by search engines. This occurs when the link is presented in a

format that is difficult or impossible for the search engine to extract from the page. The following are some examples of this, all cases where a user could easily access the link in a browser but the search engines cannot "see" the link: JavaScript; Java; Flash; Shockwave; ActiveX. All of these are examples where a programming language could be involved and so the link could theoretically be created at program run-time, meaning that even a text search of the program would not reveal it. As an extreme, but probably not uncommon, example, a site with a Java menu would have transparent navigation to typical users but could be opaque to crawlers. The server-side image map is another technique that is, in practice, impossible to index (Thelwall, 2002).

Research Design and Procedures

Different types of Web site (e.g. commercial and academic sites) have different characteristics and histories of development. The sizes of sites are also different, for instance academic sites are typically larger than commercial sites. These differences can cause the Internet Archive to cover them differently in terms of when to start covering a particular site and how often to visit the site. To improve the validity and reliability of the study, only commercial Web sites will be investigated. These sites dominate the Web numerically: in 1999 83% of Web servers contained commercial content (Lawrence & Giles, 1999). Commercial Web sites play an increasingly important role in the global economy because of the growth of e-commerce and global business competition. The overall research design is therefore to select a number of countries and to compare the coverage of their commercial Web sites in the Archive. The nations were chosen in a way to allow the investigation of linguistic factors and economic factors (e.g., developed vs. developing countries).

The methods of this study closely follow those of a previous investigation into search engines (Vaughan & Thelwall, 2003b). Firstly, four countries[2] were selected. The U. S. was chosen as a country that should be well covered due to its long history of Web use and perhaps the location of the Archive. China was chosen as a newer Internet user with a non-ASCII language, a contrast to the U. S., and Singapore was included as another relative newcomer, but with predominantly English Web pages. Taiwan was added as a more economically advanced nation with a non-ASCII language. Each of the countries has a main domain name ending for its commercial Web sites: .com (U.S.); com.tw (Taiwan); .com.cn (China); .com.sg (Singapore). A sample of Web sites was produced by randomly generating domain names beginning with www. and finishing with the above national commercial ending, but with up to four letters for the middle part of the name. To illustrate this, the sample set for Singapore was taken from the set of domain names from www.a.com.sg to www.zzzz.com.sg. This sample is not fully representative of commercial Web sites within the countries, but a pure random sample is impossible to obtain (Vaughan & Thelwall, 2003b). The main biases of selection algorithm are: long domain names are excluded, which are more likely to come from newer Web sites; and sites on shared domains are excluded, perhaps from older or poorer Web sites. The biases are likely to be relatively small within and between countries so that the comparisons among countries made in the study should not be seriously affected.

The sites identified for each country from the above process were then manually checked to ensure that only those meeting the criteria for the study would be retained. The following types were excluded (Vaughan & Thelwall, 2003b):

- Personal sites, for example sites that contained baby or wedding photographs, which are not truly commercial in nature;
- Sites that explicitly indicate that they are under construction;

[2] Taiwan is not recognized as an independent country by the United Nations and other international organizations. It is therefore more accurate to use the wording country/region here. However, for the convenience of reading and writing, the wording "country" was used throughout the paper instead "country/region".

- Sites that are selling the domain name owned;
- Sites that are password protected and not accessible to search engines; and
- Sites that are search engine interfaces themselves.

The sites were also checked for language and those in a language other than the one being investigated for the country were excluded (e.g., non-English Singapore sites were filtered out). Finally, since .com is a multinational domain, all .com sites were visited to identify the address of the owning company. Only those with addresses in the U. S. were kept. Thirty-nine of the .com sites belonged to companies from China and so they were moved to this set. The sample sizes for the U.S., China, Taiwan, and Singapore after this process were 143, 143, 141, and 94 respectively.

The Internet Archive was queried for each chosen site to see whether at least one page was indexed in it. At the same time the earliest date for the site to appear in the Archive was identified. The Vaughan & Thelwall (2003a) study repeated a set of Archive queries over a period of time and found the results to be identical. The only exceptions were when a site's owners had banned their site from the Archive. This is evidence of the robustness of results from the Archive and so queries were only submitted once.

Since unbalanced representation of different countries was found in the Archive (details in "Results" below), a technical cause of this bias was further investigated. The visibility of a Web site, as measured by the number of links from other sites to the site in question, is known to affect a site's chance of being indexed by search engines (Lawrence & Giles, 1999; Vaughan & Thelwall, 2003b), and so this could also be a factor in Archive coverage. To investigate this possibility, link count data was collected. Ideally, the Alexa search engine that feeds the Archive should be used to search for links to the Web sites (inlinks) in the study. However, Alexa does not provide a link search capability, and so link search results from other major commercial search engines were collected as a proxy. AltaVista and AllTheWeb were used (the two largest search engines that can report external inlink counts) to count the number of links to each Web site from *other* Web sites. The average link count of these two engines was used as the measure of site visibility. These figures will reflect the links indexed by the search engine rather than all Web links or Archive-indexed links. Since search engines tend to have significant overlaps (Lawrence & Giles, 1999) the figures should be reasonable estimates for the Archive-indexed links.

The imaginary site of www.abc.com will be used to explain the syntax of the query to search for external links in AltaVista and AllTheWeb. In AltaVista, the command line "link:www.abc.com AND NOT host:www.abc.com" was entered into the query window of the advanced search mode. In AllTheWeb's advanced search mode, "www.abc.com" was entered into the window in between the windows of "must include" and "in the link to URL" under "Word Filters". Further, "www.abc.com" was entered in the "exclude" window under "Domain Filters."

An additional issue is the reliability of results from search engines. Most search engines give an estimate for the total number of pages matching a query in addition to returning links to the first few. These estimates are known to be potentially unreliable (Bar-Ilan, 1999; Mettrop & Niewenhuysen, 2001; Rousseau, 1999; Snyder & Rosenbaum, 1999). However, recent studies have shown that search engines have become much more stable (Thelwall, 2001; Vaughan & Thelwall, 2003a; Vaughan & Hysen, 2002; Vaughan, forthcoming), making this issue less of a concern at the time of the study. Nevertheless, to improve the reliability of data collected, two rounds of data were collected for the same link search query with about three weeks in-between the rounds. Results from both search engines were found to be stable in that the two rounds of data are highly correlated (the Spearman correlation coefficient was 0.999 and 0.992 for AltaVista and AllTheWeb respectively. In fact, the two rounds of AltaVista link counts were identical for 90% of sites). The average of the two rounds of data was taken and used in all the statistical analyses involving this variable.

Results

The number of Web sites that were in the Archive is tabulated by country in Tables 1 and 2. A chi-square test on the table shows a significant difference ($p < 0.01$) among different countries in their inclusion by the Archive. Comparing the two counts, the U. S. sites are over represented in the Archive (the observed count is much higher than the expected count for the "in Archive" cell) while China is under represented. Taiwan and Singapore are fairly proportionally represented.

Table 1. Country representation in the Archive.
Note: The main number is the number of sites found and the number in brackets is the expected count.

Country	In Archive	Not in archive	Total
China	83 (104.8)	60 (38.2)	143
Singapore	66(68.9)	28 (25.1)	94
Taiwan	102 (103.4)	39 (37.6)	141
U.S.	131 (104.8)	12 (38.2)	143
Total	382	139	521

Table 2 Country representation in the Archive by percentage

Country	In Archive	Not in archive
China	58%	42%
Singapore	70%	30%
Taiwan	72%	28%
U.S.	92%	8%
Average	73%	27%

There is an unbalanced coverage of different countries in the Archive. What could be the possible reasons for this bias? Two technical causes, the language of the site and the visibility of the site, were investigated. The low Archive inclusion of sites from China could not be attributed to the language factor. Taiwanese sites (in the Chinese language) are better represented in the Archive than sites from China. In fact, the percentage of Taiwanese sites in the Archive (72%) was slightly higher than that for the English language Singapore sites (70%).

Could Inlink Counts be a Factor in Archive Inclusion?

Inlink count data were collected using the commercial search engines AllTheWeb and AltaVista (see "Methodology"). All sites in the study were cross-tabulated by their presence in the Archive and whether they have links to them or not, as shown in Table 3. A chi-square test on the Table 3 data shows a significant relationship ($p < 0.01$) between a site's presence in the Archive and whether there is a link to the site. Again, both the observed count and the expected count are presented in each cell of the table. Comparing the two counts, Web sites that have links to them are more likely to be included in the Archive. About 40% (207/521) of sites have no links to them (i.e., the two search engines both retrieved zero hits). Among them, only 46% (95/207) were included in the archive. In contrast, 91% (287/314) of the sites that have at least one link to them are included in the Archive.

Table 3 The relationship between Archive presence and links to a site
Note: The number in brackets is the expected count

	Has links	Has no links	Total
In Archive	287 (230.2)	95 (151.8)	382
Not in Archive	27 (83.8)	112 (55.2)	139
Total	314	207	521

Can the difference among different countries' Archive presences all be attributed to the difference in their Web site visibility, i.e. site inlinks? A two-way analysis of variance (ANOVA) test was carried out to examine this question. The dependent variable of the test is the inlink count and the two independent variables are country and Archive presence. The frequency distribution of the inlink count is very skewed, which is to be expected given the common occurrence of power law phenomena on the Web (Broder et al., 2000; Rousseau, 1997; Thelwall & Wilkinson, 2003). However, this violates the normality requirement of the ANOVA test. A log transformation (Howell, 2002, pp. 342-349; Judd & McClelland, 1989, 493-528) was applied to link count data and the ANOVA test carried out on the transformed data. The test result shows a significant difference ($p < 0.01$) between inlink counts among the four countries and between the sites covered by the Archive and those not covered (see Figure 1).

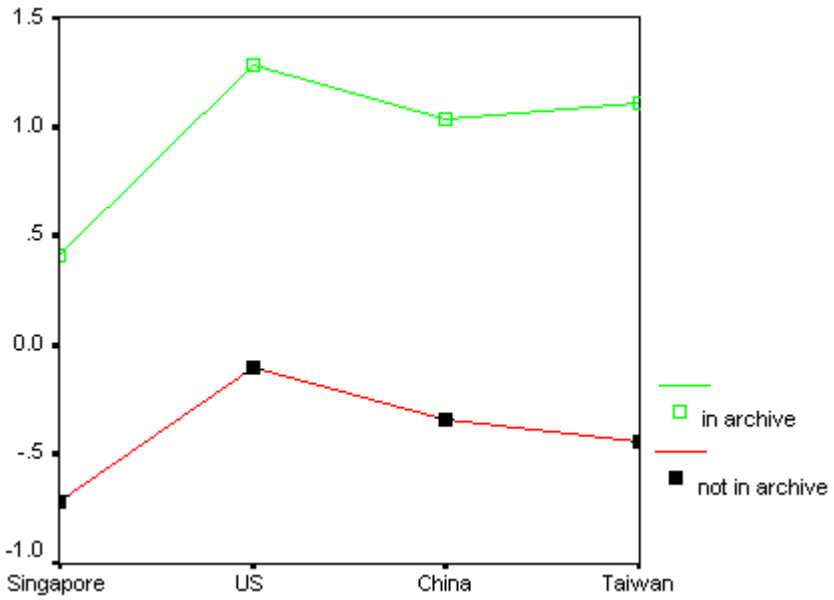

Figure 1. Inlink count comparisons

The vertical axis of Figure 1 represents the logarithm of the inlink counts and the horizontal axis represents the four countries. The two lines on the figure represent the two groups of Web sites in the study, those included in the Archive and those not included. The line that represents the group included in the Archive sits much higher, which means that Web sites that are included in the Archive have much higher average inlink counts. The chi-square test used binary data (whether a site has an inlink or not) while the ANOVA test used the actual number of links; both point to the same conclusion.

Figure 1 shows the significant differences among the countries. The average inlink counts for the U. S. sites are higher than the others. The two lines in Figure 1 are close to parallel, which shows that there is no significant interaction between the two independent variables of the ANVOA test. This means that for both cohorts of sites (those in the Archive and those not), the difference among the countries is very similar: U.S. sites have the highest inlink counts on average. Singapore is the lowest while China and Taiwan are similar and lie in the middle.

Archive Age

For all sites that were included in the Archive, the date of their first appearance in the Archive was recorded. These dates were used to calculate the Archive Age of the sites (how long the Archive had been covering the sites) expressed as the number of months at the end of the year 2002. The Archive Age variable followed a normal distribution. When sites from all countries were combined, the average age is about 36 months (i.e. on average these sites have been in the archive for three years). Is there a

difference among different countries in their Archive Age? A one-way analysis of variance test shows a significant difference. Table 4 shows the average age by country. A Tukey's HSD test shows three cohorts: the U. S. sites; Taiwanese sites; sites from China and Singapore with no significant difference between the two. The U. S. sites have been in the Archive for four years on average, almost twice that of sites from China. On the other hand, the U. S. sites' Archive ages also have a larger variability as measured by the standard deviation. Although China sites are slightly younger than those from Singapore, this difference does not seem to be large enough to account for the difference in coverage.

Table 4. Archive age comparisons
(Age is measured as the number of months at the end of 2002)

Country	Average Archive Age	Standard Deviation
China	25	8
Singapore	28	11
Taiwan	34	16
U.S.	48	18

Could Inlink Counts be a Factor in Archive Age?

Could the visibility of a site, as measured by the number of inlinks to the site, affect how long the site has been covered by the Archive? A Spearman correlation coefficient test was used to test this hypothesis and a significant relationship was found between inlink counts and Archive age. When data from all countries were combined, the correlation coefficient was 0.59. When the correlation was calculated separately for different countries, the coefficients were 0.66, 0.62, 0.66, and 0.49 for China, Singapore, Taiwan, and the U.S. respectively. All coefficients reported here are highly significant ($p < 0.001$).

Archive Age Analysis with Links Taken into Consideration

Since significant differences among different countries exist in both the Archive age and inlink counts (the two-way ANOVA result reported earlier), and a significant relationship between inlink counts and Archive age was also found, it is logical to analyze Archive age whilst taking inlink counts into consideration. This was attempted through the creation of a new variable, the age-link ratio. The ratio was calculated for each site by dividing the Archive age by the inlink count, but only for sites with at least one inlink. This measures the number of months a site has been in the archive per inlink it had received. The frequency distribution of the age-link ratio was very skewed because the link distribution was skewed as discussed above. A logarithm transformation was again applied and a one-way analysis of variance (ANOVA) test carried out afterward. The ANOVA test and the follow up Tukey's HSD test show that there is no significant difference in the age-link ratio for Web sites from the U.S., China, and Taiwan. However, the age-link ratios for Singapore sites were on average higher than those of the other three countries. This means that there is no significant difference among China, U. S., and Taiwan sites in how long the Archive has been covering them once inlink counts have been factored out. In other words, the relatively long coverage of the U. S. sites can be explained by those sites being more visible on the Web (i.e., having more inlinks to them). Although Singapore sites have a shorter history of being in the Archive compared to the U.S sites, the former has actually been in the Archive relatively longer given the very low inlink counts of those sites. The analysis has been unable to explain Singapore's coverage through the link count and Archive age data.

Discussion

There is uneven representation of different countries in the archive. The average percent of sites in the Archive for the four countries is about 73%. The percentage of U.S. sites in the Archive (92%) is in strong contrast to that of China (58%). The percentages for Singapore and Taiwan are similar and close to the average of 73%. The differences seem to be too large to be accounted for by cultural differences that would affect the domain name sampling technique used. This unevenness is not attributable to difficulties in indexing non-ASCII languages, but attributable to the number of links to a site. Given that crawlers find new sites by following links from previously indexed sites, this is a logical conclusion and was backed up by the data collected. Commercial search engine coverage of sites is also biased by inlink counts (Lawrence & Giles, 1999; Vaughan & Thelwall, 2003b).

The average Archive age is about three years. There is a significant difference among different countries. The U.S. has the longest average age, followed by Taiwan. China and Singapore have the shortest history, and there is no significant difference between the two. Again, Web site visibility as measured by links to the site is a potentially explanatory factor. Inlink counts correlate with the Archive ages of sites. This relationship needs further analysis, however, to determine the possible causes of the correlation. First, it is unlikely that Archive inclusion generates inlinks since the Archive does not have as high traffic as search engines and its primary use is not to find new pages with useful content. Higher search engine indexed inlink counts could make it more likely for a site to be archived earlier, however, since a site with more inlinks at any given moment in time would be more likely to be found by the archive spider. Based upon the network growth model (Albert & Barabasi, 1999) and other research results (Vaughan & Thelwall, 2003a; Vaughan & Thelwall, 2003b), a site with more inlinks is likely to be older and therefore has a chance of being indexed earlier by the Archive spider than a newer site. As a result, higher inlink counts are likely to associate with older and more easily found sites. Logically, then, if different countries have a different spread of this type of site then different Archive coverage of them could be expected. The final test showed that this could indeed explain the differing rates for three of the four countries, the exception being Singapore, which seems surprisingly well indexed given its low inlink counts.

There are possible explanations for the case of Singapore. The age and inlink based model used in this paper is imperfect because inlinks probably represent a combination of age and site quality (Bianconi & Barabasi, 2001). So a country with a preponderance of new but better quality sites, rather than a spread of combinations of the two, would not fit the model. Since the link influence model was only an estimate, firm conclusions cannot be made from it that links are not a possible explanation for countries that do not fit the pattern, only that links are a potential explanation for countries that do fit it. As a result, since three of the four countries fit the same general pattern, it is not reasonable to interpret the findings as evidence of a systematic positive bias in Archive indexing towards Singapore.

The choice of countries for the study does not cause a problem for the main findings because significant bias has been found. If no bias had been discovered then further research would have been necessary to discover whether the lack of bias was universal. The findings are limited, however, only providing an indication of the *extent* of bias between the four countries in the set. For example, it would be of interest to know how coverage of other richer countries, such as those in Europe, would compare to that of the U. S. Additionally, it would also be useful to know if there are factors other than links and age that influence Archive coverage.

In summary, although there is significant bias in terms of both rates of inclusion in the Archive and length of time of inclusion by country, both phenomena are broadly consistent with being the natural results of the evolving nature of the Web and its evolving link structure. In other words the Internet Archive is naturally biased by *link structures* rather than by *countries*, but historical factors have caused the first to map onto the second.

Conclusion

The unbalanced representation of different countries in the Internet Archive, although not intentional, remains a problem in providing an effective historical archive of the Web. It is reassuring, however, that the Archive does not seem to have bias against indexing of non-ASCII pages, although it seems likely from the discussion that poorer countries generally will be under-represented. If the Archive truly wants to be a mirror of Web history, measures need to be taken to correct the imbalance that currently exists. However, this may not be a practical possibility since sites can only really be found in large numbers by following links, unless domain name allocating organizations would be willing to give lists of used names to the Archive.

Archive bias is probably not a significant problem for one type of scholarly user: those chasing missing URLs cited in journal articles. The existing national and linguistic biases in scholarly communication make this additional factor relatively unimportant. For longitudinal or historical studies of a type that are international in character or specifically compare countries (e.g., Leydesdorff & Curran, 2000) the bias provides a real obstacle. The problem is great enough to rule out both types of retrospective comparative studies based upon the archive unless methods are found to avoid the bias. For instance if only a small group of countries are studies then the above method can be used to assess bias within the sample, and steps taken to compensate. Inevitably, for some research questions this will not be possible and so they will simply not be answerable. This will particularly affect studies related to site age and link structures, the apparent causes of the biases found.

For researchers conducting longitudinal or historical studies within a single nation, or worldwide studies without any interest in international dimensions, although the findings do not specifically address their concerns it is reasonable to believe that there will also be intra-national and other biases that are related to site age and link structures. Caution must therefore be advised in interpreting findings of such studies unless methods can be devised to bypass these problems.

The discovery that the only serious resource for large quantities of historical Web data is flawed is a setback for Web research.

References

Albert, R. & Barabasi, A. (1999). Emergence of scaling in random networks, *Science*, 286, 509-512.

Albitz, P. & Liu, C. (2001). *DNS and BIND (4th ed.)*. Sebastopol, California: O'Reilly.

Bar-Ilan, J. (1999). Search engine results over time: A case study on search engine stability. *Cybermetrics*, 2/3. Available at: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html. (Accessed July 30, 2002).

Bianconi, G. & Barabasi, A. (2001). Competition and multiscaling in evolving networks, *Europhysics Letters*, 54, 336-422.

Broder, A., Kumar, R, Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph structure in the Web, *Journal of Computer Networks*, 33, 309-320.

Brown, R. H., & Davis-Brown, B. (1998). The making of memory: the politics of archives, libraries and museums in the construction of national consciousness, *History of the Human Sciences*, 11, 17-32.

Burner, M. (1997). Crawling towards eternity: Building an archive of the World Wide Web. *Web Techniques*, 2. Available: http://www.newarchitectmag.com/archives/1997/05/burner/. (Accessed April 30, 2003).

Chavez-Demoulin, V.C., Roehrl, A.S.A., Roehrl, R.A., Weinberg, A., (2000): The WEB archives: A time machine in your pocket, Internet Archive Colloquium, San Francisco, March 2000, Available: http://citeseer.nj.nec.com/chavez-demoulin99web.html. (Accessed April 30, 2003).

Council on Library and Information Resources. (2002). Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving. Washington D.C.: Council on Library and Information Resources and the Library of Congress. Available: http://www.clir.org/pubs/reports/pub106/contents.html. (Accessed 28 November, 2002).

Derrida, J. (1996). Archive fever: A Freudian impression. Chicago: Chicago University Press.

Eagleton, T. (1996). Literary theory: An introduction (2nd Ed).Oxford: Blackwell.

Ernst, W. (1999). Archival action: the archive as ROM and its political instrumentalization under National Socialism. *History of the Human Sciences*, 12, 13-34.

Featherstone, M. (2000). Archiving cultures. *British Journal of Sociology*, 51, 161-184.

Fox, E. A., & Urs, S. R. (2002). Digital libraries, *Annual Review of Information Science and Technology*, 36, 503-589.

Green, H. (2002). A library as big as the world. BusinessWeek.com. Available: http://www.businessweek.com/technology/content/feb2002/tc20020228_1080.htm. (Accessed 28 November, 2002).

Hawking, D., Craswell, N. & Thistlewaite. P. (1999). Overview of the TREC-7 very large collection track. In E.M. Voorhees & D.K. Harman (Eds), Proceedings of the Seventh Text REtrieval Conference (TREC-7) (pp. 91-104). Gaithersburg, Maryland: NIST Special Publication 500-242.

Howell D. (2002). Statistical Methods for Psychology, 5th ed., Pacific Grove, CA: Duxbury.

Internet Archive (2002a). About the Internet Archive. Available: http://www.archive.org/about/about.php. (Accessed November 27, 2002).

Internet Archive (2002b). News Forum. Available: http://www.archive.org/iathreads/forum-display.php?forum=news. (Accessed 4 December, 2002).

Judd, C. M., & McClelland, G. H. (1989). Data analysis: A model-comparison approach, San Diego: Harcourt Brace Jovanovich.

Ju-Pak, K. H. (1999). Content dimensions of web advertising: A cross-national comparison. *International Journal of Advertising*, 18, 207-231

Kenney, A. R., McGovern, N. Y., Botticelli, P., Entlich, R., Lagoze, C., & Payette S. (2002). Preservation risk management for Web resources: Virtual remote control in Cornell's project Prism. *D-Lib Magazine*, 8. Available: http://www.dlib.org/dlib/january02/kenney/01kenney.html. (Accessed 28 November, 2002).

Koehler, W. C. (1999). An analysis of web pages and web site constancy and permanence. *Journal of the American Society of Information Science*, 50 (2), 162-180.

Koman, R. (2002). How the Wayback Machine Works, Available: http://www.oreillynet.com/pub/a/webservices/2002/01/18/brewster.html. (Accessed 7 February, 2002).

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. Nature, 400, 107-109.

Leydesdorff, L. (2001). Indicators of innovation in a knowledge-based economy, *Cybermetrics*, 5. Available: http://www.cindoc.csic.es/cybermetrics/articles/v5i1p2.html. (Accessed May 1, 2003).

Leydesdorff, L. & Curran, M., (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy, *Cybermetrics*, 4. Available: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html. (Accessed May 1, 2003).

Lyman, P. (2002). Archiving the World Wide Web, School of Information Management and Systems, University of California, Berkeley. Available: http://www.clir.org/pubs/reports/pub106/web.html. (Accessed 28 November, 2002).

Masanès, J. (2002). Towards Continuous Web Archiving: First Results and an Agenda for the Future, *D-Lib Magazine*, 9. Available: http://www.dlib.org/dlib/december02/masanes/12masanes.html. (Accessed 18 December, 2002).

McMillan, S. J. (2001). Survival of the fittest online: A longitudinal study of health-related Web sites. *Journal of Computer Mediated Communication* 6(3). Available: http://www.ascusc.org/jcmc/vol6/issue3/mcmillan.html. (Accessed May 1, 2003).

Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines - fluctuations in document accessibility. *Journal of Documentation*, 57, 623-651.

OCLC Web Characterization Project (2002). Country and Language. Available: http://wcp.oclc.org/. (Accessed: September 4, 2002).

Osborne, T. (1999). The ordinariness of the archive. *History of the Human Sciences*, 12(2), 51-64.

Rauber, A., Bruckner, R. M. Aschenbrenner A., Witvoet, O. & Kaiser, M. (2002). Uncovering information hidden in Web archives: A glimpse at Web analysis building on data warehouses, *D-Lib Magazine*, 9. Available: http://www.dlib.org/dlib/december02/rauber/12rauber.html, accessed 18 December, 2002.

Rousseau, R., (1997). Sitations: an exploratory study, Cybermetrics, 1. Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3. Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html. (Accessed July 30, 2002).

Snyder, H., & Rosenbaum, H. (1999). Can search engines be used as tools for web-link analysis? A critical view. *Journal of Documentation*, 55, 375-384.

Suel, T. & Yuan, J. (2001). Compressing the graph structure of the web. In Proceedings of the IEEE Data Compression Conference (DCC), march 2001. Available: http://citeseer.nj.nec.com/cache/papers/cs/23869/http:zSzzSzcis.poly.eduzSzsuelzSzpaperszSzgraph.pdf/suel01compressing.pdf accessed 28 November, 2002.

Thelwall, M. (2000), Commercial Web sites: Lost in cyberspace?, *Internet Research*, 10, 150-159.

Thelwall, M. (2001), The responsiveness of search engine indexes, *Cybermetrics*, 5, http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html.

Thelwall, M. (2002). Methodologies for crawler based Web surveys, *Internet Research: Electronic Networking and Applications*, 12, 124-138.

Thelwall, M. & Wilkinson, D. (2003). Graph structure in three national academic Webs: Power laws with anomalies, *Journal of the American Society for Information Science and Technology*, 54(8), 706-712.

Uberti, T. E. (2003). Commercial, Technological and Information Flows: some elements to analyse the globalisation process. (PhD thesis). Milan, Italy: Universita' Cattolica del Sacro Cuore.

Vaughan, L. (forthcoming). New measurements for search engine evaluation proposed and tested. To appear in *Information Processing & Management*.

Vaughan, L. & Hysen, K. (2002). Relationship between links to journal Web sites and Impact Factors, *Aslib Proceedings: New Information Perspectives*, 54, 356-361.

Vaughan, L. & Thelwall, M. (2003a). Scholarly use of the Web: What are the key inducers of links to journal Web sites? *Journal of the American Society for Information Science and Technology*, 54, 29-38.

Vaughan, L. & Thelwall, M. (2003b, to appear). Search engine coverage bias: evidence and possible causes. *Information Processing & Management*.

Wilson, L. (2001). One on one with Brewster Kahle, co-founder of Internet Archive. Available: http://sanfrancisco.bizjournals.com/sanfrancisco/stories/2001/10/22/newscolumn9.html. (Accessed 28 November, 2002).