

Commercial Web sites: lost in cyberspace?

Item Type	Journal article
Authors	Thelwall, Mike
Citation	Thelwall, M. (2000), "Commercial Web sites: lost in cyberspace?", Internet Research, Vol. 10 No. 2, pp. 150-159. https://doi.org/10.1108/10662240010322939
DOI	10.1108/10662240010322939
Publisher	MCB UP Ltd
Journal	Internet Research
Download date	2026-05-17 22:09:32
Link to Item	http://hdl.handle.net/2436/4505

Commercial Web Sites: Lost in Cyberspace?

Internet Research Volume 10 Number 2, 2000, pp 150-159.

Mike Thelwall

School of Computing and Information Technology
University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1SB, UK.

Email: cm1993@wlv.ac.uk

Abstract

How easy are business web sites for potential customers to find? This paper reports on a survey of 60087 web sites from 42 of the major general and commercial domains around the world to extract statistics about their design and rate of search engine registration. Search engines are used by the majority of web surfers to find information on the web. However, 23% of business web sites in the survey were not registered at all in the five major search engines tested and 82% were not registered in at least one, missing a sizeable potential audience. There are some simple steps that should also be taken to help a web site to be indexed properly in search engines, primarily the use of HTML META tags for indexing, but only about a third of the site home pages in the survey used them. Wide national variations were found for both indexing and META tag inclusion.

Keywords

Search engine, World Wide Web, Business

Introduction

The commercial potential of the Web is a subject of widespread discussion, with many predictions of continuing rapid growth in the future. Of great concern, therefore, is the question of how to produce a successful Web intervention for a business. One useful model summarising the findings of previous research is Simeon's (1999) attracting, informing, positioning and delivering (AIPD) approach to evaluating web sites. The first hurdle for any web site, however, is the first aspect: attracting visitors. If the site does not get visited then its content is irrelevant. A major problem is the size of the web. According to Lawrence and Giles (1999), there were approximately 800 million publicly indexable pages in February 1999, with 83% of them coming from commercial sources. One of the attractions of the Internet is the size of its user base and the fact that information can be accessed by any of them at a low cost. But publishing a website is very different from placing an advert or shop by the side of a busy road: those travelling on the information superhighway will not 'notice it whilst passing by' unless additional steps are taken to make it visible. Although users may be able to correctly guess the address of *known* companies, a better analogy would be buying an office phone: unless the number is publicised or lodged in phone books it is unlikely to be called. Publicising a web site is increasingly being attempted through traditional advertising (Pardun and Lamb, 1999), but can also be achieved through including the address on all office stationary for existing customers or placing online adverts on other sites. The equivalent of placing an entry in a phone book is registering the site with search engines and directory web sites and

attempting to get it linked to by relevant gateway sites. Most web users employ search engines as part of at least one of their strategies to find new web sites (CyberAtlas, 1999). Any web site owner that wants to attract new visitors should therefore be concerned whether or not their site is registered in the major search engines. But search engines do not cover the whole of the web, only around 16% at most (Lawrence and Giles, 1999), and so many must be left out.

Much research and development time goes into designing, analysing and assessing search engines and methods of indexing or classifying web pages and sites (Brin and L. Page, 1998; Chun, 1999; Dowe *et al.*, 1998; Gordon and Patak, 1999; Henzinger, *et al.*, 1999; Kirsch, 1998; Pringle *et al.*, 1998; Schwartz, 1998; Snyder and Rosenbaum, 1998; Spink *et al.*, 1999). Search engines typically use two methods to find web pages to index: by following links from previously registered sites and by allowing users to register the addresses of unknown sites. If a page is not registered in a search engine there are two possible reasons: because it has not found the site or because it found the site but has decided not to index it. A web site designer has full control over the first of these options because search engine registration is easy and usually free. He or she also has partial control over the latter, as discussed later. An important question is, therefore, how seriously web designers are taking search engine coverage. It is difficult to ask this question directly of the designers because it is essentially asking them whether they have exhibited shortcomings in their job, and so a questionnaire on this topic would be expected to be extremely unreliable. It is, however, possible to tell from web sites whether they have been designed to be 'search engine friendly' or not.

Internet search engines are an important information resource on the Internet, and if the web begins to take a significant share in commerce, then any differences in national coverage can have major financial implications. Such differences may stem from differing national use of the Internet, from different treatment by search engines or from the linguistic implications on search algorithms for countries with languages differing from the dominant American English. Indeed there are now many language-specific and country-specific search engines as well as the multilingual, general type.

In this paper a survey of a large number of web sites in the largest predominantly commercial web domains is discussed, to ascertain the extent of search engine coverage in each as well as the use of key design features, and any national differences between them.

Research Design

A predominantly automated survey was conducted on 42 of the largest general or commercial domains of the Internet. The first task was to choose which domains to cover. Commercial domains were chosen over other types because almost certainly the vast majority are built with the hope of attracting new business. For example, in the UK a survey indicated that only 1% of the co.uk domain appeared to be intended for a closed audience (Thelwall, 1999). In contrast, the second largest Internet application, education, has a much more complex set of purposes for web sites (Middleton, *et al.*, 1999) and the same is believed to be true about the other domains.

The largest 40 national domains were selected by using the Yahoo!Directory listings as an approximate indicator of relative domain size. India was also included as a very large country which would have otherwise been omitted. The large general com domain completed the set of 42.

National domain name structures vary so that there is not one common place to search for commercial sites. Most European national domains aggregate together

all applications in a single domain, such as .fr for France. Many other countries have a specific commercial sub-domain, com or co, for example .co.il for Israel and .com.eg for Egypt. There may also be more than one commercial sub-domain, for example, in the UK the choices include .co.uk, .plc.uk and .uk.com, although the first appears to be by far the most common. The apparent neat splitting of web sites along easily identifiable national domain lines is complicated by the fact that the .com domain can also be used by companies from any country.

For each country, if a well-used commercial sub-domain was identified then this was used in place of the general national domain. It is expected that even the general national domains would be dominated by commercial sites because 83% of web pages are of commercial origin, as mentioned previously.

A method of selecting a random sample of sites from each domain was needed. In previous surveys random selections have been made from search engine directory listings. This method is unsatisfactory for this study because it would restrict the sample to sites registered in a search engine, compromising the objectives. The alternative strategy adopted was to design a web crawler that tested legal domain names to see whether a site with the name existed. A truly random survey of this type would select at random a domain name from a given domain, for example in the domain .com, it might choose `www.a-xy56zo6qrst2xaefgb0ji7rw.com` as one potential name. It would continue randomly generating and testing names until the necessary sample size for real domains had been reached. Unfortunately the huge number of legal names (22^{37} for the com domain) means that with statistical near certainty no used names would be found by a computer using this method for any realistic length of time. An alternative to this impossible ideal solution was therefore adopted, which was to survey only a subset of each domain or sub-domain. For practical purposes small names were chosen and all domain name parts of up to three letters were surveyed. For example in the com domain the names `www.a.com` to `www.zzz.com` were tested to see whether they existed. This is clearly not a genuinely random choice and will have inherent biases, including perhaps towards longer established sites that may well have shorter domain names, a factor particularly in more densely subscribed domains such as .com. Since this is unable to be a random sample and it was easy to automatically survey a large number of sites, the whole three-letter domain area was surveyed instead of just finding a fixed size sample in each case. Once the sites had been identified, they were interrogated by a crawler program to record basic statistics on site design. The statistics were culled from the home page and, in the case of a frames-based page, all sub-frames. The crawler program also tested to see whether the sites were registered in a range of popular search engines. These search engines were chosen for size of user base and their ability to allow a reliable search for a specific site using its domain name.

The survey was executed from 19 July 1999 to 30 July 1999 on software running on 72 computers connected to the UK academic section of the Internet through an ATM connection. Each site was given a time limit of sixty seconds in which to download the HTML of its main page, without images or other resources of any kind. This was another practical limitation that again imposed some bias on the sample taken, this time in favour of sites with faster or less busy connections. Web sites on servers that were down would also be missed by this method.

For each site downloaded a check was made to see whether it included JavaScript, Java, Frames and the keywords and description META tags. In the case of a frameset page, the Java and JavaScript checks also include all embedded frames. The site was then checked for registration in a number of the most used search

engines that allow direct checking of URLs. Search engines reporting a busy message were checked again later. The search engine tests merely tested whether any pages at all in the site were registered rather than the proportion that were registered. It would be expected that in most cases when some pages in a site were registered, not all were. The process of indexing sites takes time, the duration depending on the search engine and which method is eventually used to find the page. It is also known that search engines do not normally index a site systematically but have complex algorithms to decide in which order to follow all new links submitted and found, as well as to check up on previously indexed pages (Tunender and Ervin, 1998) which may lead to significant time delays before indexing pages deeper in a site. Some search engines also have maximum depth policies for sites, following only a set number of levels of links from a submitted starting page. Furthermore, a web site designer may expect users to be able to find a web site through a single entry page, or might want it to be able to be accessed through any of the pages. The extent of search engine coverage of a site is therefore not necessarily an issue for all sites. However, if a site is not indexed at all in a search engine then it clearly cannot be found from it. The test of the existence of any pages in a search engine is therefore the most appropriate one.

Once all data had been gathered a manual check was performed to remove spurious sites. These included sites returning error messages, empty directory listings, access forbidden messages, server default pages, under construction pages and various types of domain name holding pages. Further checks were made to remove pages that occurred under multiple domain names due to server redirection for example. No attempt was made to deal with sites occurring in multiple domains, for example the redirection of a com site to a co.uk address. Such sites were counted in both domains.

Results and Discussion

Search Engine Coverage

Table I shows the search engine coverage results, and is arranged in order of AltaVista coverage. This was the search engine tested that indexed the most sites. The figures for search engine coverage of up to 64% are much higher than the maximum of 16% (15.5% for AltaVista) from February 1999 deduced by the Lawrence and Giles (1999) study of a genuinely random sample of the entire web, following up an earlier, higher estimate (Lawrence and Giles, 1998). This relatively large figure is expected because it indicates sites with at least one page indexed in the search engine, rather than a result for individual pages.

Take in Table I

The overall results show far from universal search engine coverage. At least 63% of every domain appears to be completely omitted from at least one major search engine. A majority of sites will therefore be unable to be found by a significant minority of surfers. Nearly a quarter of sites surveyed are also not registered in any of the search engines that were tested, effectively invisible to a large majority of search engine users. This should be a serious cause for concern for the site owners even if they are using other means to attract visitors.

There are two clear intermingled general trends in the data: for more economically advanced countries to register better and for general national domains to perform better than commercial ones. From the figures of Lawrence and Giles the national domains may well be approximately 83% commercial and so it is believed that the economic trend is the significant one. An indicator of this is that the five

countries in the survey that have less than one web site per million of population in the area searched come in the bottom 8 of the table. This could be a result of better practice in these countries, a longer pedigree embedding sites better in the databases or perhaps structural or deliberate partiality in the search engine software. A further confusing factor is the language issue. Web pages in non-English speaking countries are not surprisingly often in the local language. But since English is still the dominant language of the web this may lead to inefficient indexing of web pages or a downplaying of their importance due to the relative infrequency of demand through the mainly English web searches. This does not seem to have happened, with English speaking national domains not scoring particularly well, although Infoseek does seem to score them better relative to the other search engines. A further complication comes from Unicode pages used in countries with non-ASCII based languages. This does not seem to have had a large impact on the results however, with Japan scoring well.

There are a number of unusual features in the table that will be mentioned following a more detailed discussion of search engine indexing practice. This is an extremely important issue for web page designers, and one which has spawned many books and web sites, including the excellent Northern Webs' set of pages (1999). An important consideration in addition to whether a search engine will find a page is whether it will be indexed once found or whether it will be discarded (then or subsequently) because it is perceived to be trying to mislead the page ranking algorithm or is judged not sufficiently popular or relevant (Kirsch, 1998; Laursen, 1998; Marchiori, 1997; Tunender and Ervin, 1998). One quoted measure of popularity is the number of links to a page from other pages. This has led to the phenomena of links exchanges where a group of sites agree to link to each other in order to artificially boost their popularity ratings.

A domain could score well with search engines by its sites excelling at being registered or being retained in the database for a number of reasons. A major factor could be web design companies, ISPs and domain name organisations offering cheap multiple search engine registration packages. It could also do well perhaps as a result of government policy in financing gateway sites or supporting web education initiatives, or due to a generally high level of understanding amongst web page designers, or for accidental reasons. For example, relatively small national domains with well organised gateway sites linking to the majority of national sites could expect to do well in getting their sites registered. These gateway sites would ensure that most pages got through the first hurdle of being found by search engines and would have the popularity advantage of already being pointed to. This may be a reason for Finland scoring well with AltaVista and extremely well with MSN, 11% better than the second placed com domain with the latter, with gateway sites such as the Companies Information Gateway (helecon.hkkk.fi/ENTERPR/) which links to the home page of over 270 companies and groups of companies perhaps making a real difference. This particular site is registered well in AltaVista and even InfoSeek, but the latter had not (at the time of checking) yet indexed all sites linked to by this site, although it does not claim to index deeply (Northern Webs, 1999). Another possible reason for Finland performing relatively poorly in Infoseek may be its high use of frames and the negative link between frames based home pages and registration discussed earlier, but this would not account for more than 4% of the Infoseek score in this case.

The age and maturity of the domains in the top half of the table may explain their relative success. This could be attributed in part to the cumulative effect of more use of the indexing META tags, more submissions to search engines in recognition of

their importance, more gateway sites and more links exchanges. The Swiss domain is a very high rating large national domain with an interesting national characteristic. Many of the commercial sites were actually produced in triplicate with three different domain names for three different languages. The links between these pages would artificially inflate their popularity rankings, a possible partial explanation for the good national performance.

Various features in the table stand out as unusual. Firstly InfoSeek performs better on the com and com.au domains than on the others, and in fact it performs quite poorly on most of the other domains, although it does relatively reasonably on all the other English speaking domains. Some national results also stand out. Japan performs well only in AltaVista, as does Russia. Indonesia is another anomaly, performing well in Yahoo, HotBot and MSN relative to AltaVista. There may be national biases in search engines for various reasons such as the creation of national versions or home pages for the search sites. The table certainly does not reflect this directly as a clear pattern. It may be that more subtle workings of the search engines account for the peculiarities found.

Two domains that might have been expected to perform better are the German and the UK commercial domains. Both of these countries are amongst the biggest users of the Internet outside the USA and have a large number of sites, yet their performance is below average. A possible explanation is one of combined and uneven development: the size of the online community and popularity of the Internet in these countries may have created an impetus that has persuaded companies with little Internet experience to have a site, resulting in a dilution of the quality of the national web. If this were true then the effect would be expected to ripple through to other countries as similar scale Internet use arrives.

Document Design

Table II shows the basic statistics on the number of sites discovered in each domain from the 18278 surveyed as well as selected details of the underlying HTML. A total of 42 domains were surveyed, 767,676 attempts to find domain names, from which a total of 76,812 were found. Of these 78% (60087) were adjudged to be genuine working web sites. The relatively high percentage of spurious pages for the com domain, 46%, mainly reflects the high number of holding pages for domain names, many of which appear to have been bought wholesale as speculative investments by a small number of companies. This may also reflect the ease and cheapness of reserving a com domain name, which can be achieved in a few minutes online with a credit card.

Take in Table II

The keywords and description META tags are optional components of a web page that are invisible when the page is viewed in a web browser but have been included in the HTML specification from version 2.0 in order to facilitate indexing of the pages by automatic processes such as search engine robots. The French national domain was the only one to show a majority of pages using these tags. The overall percentage use of these tags is very poor at 33% and 35%, slightly lower than the figure of 34.2% for either on all server home pages from February, 1999 (Lawrence and Giles, 1999). The reason for the high rate of omission is probably a combination of factors. Their invisibility mitigates against their use by inexperienced designers, but this is exacerbated by general uncertainty as to their exact impact upon search engines, which need to keep secret the fine details of their page ranking algorithms. Some

search engines such as Northern Lights ignore META tags altogether, possibly as a result of misuse in the past as part of a page designer's strategy to improve page rankings, but many do use them as the main element or an important component of their page ranking strategy (Dowe *et al.*, 1998; Kirsch, 1998; Pringle *et al.*, 1998; Tunender and Ervin 1998). A related issue is the use of an appropriate title for the home page of a site. A very large number of sites had no title or an uninformative title such as variations of 'home page', 'index' (but not a default index page), 'Page x', 'Untitled document' and 'Put your page title here'. Titles are known to be very important for the ranking and indexing of pages in a number of search engines (Pringle *et al.*, 1998; Tunender and Ervin, 1998) in addition to being the default description of a bookmark if a user chooses to make one. There is clear evidence here of widespread poor design practice. A well-constructed page would have a title and META tags appropriate for the kind of user desired at the site.

The significant proportion of sites using frames for the home page is a further cause for concern from a search engine point of view. The disadvantage of a design using frames is that search engine rankings can only be maximised with a non-frames based page. Moreover, only the outer frameset document can be reliably pointed to. The individual frames can be pointed to, but unless an intelligent server automatically redirects the request, the frame will be displayed as a frame on its own, orphaned from its frameset. Some such pages are designed to function on their own, but many are not, commonly not including any navigation aids. As a result frames pages are ignored or partially ignored by many search engines or potentially inappropriately pointed to by those that do index them. In the best case where a frames and a non-frames version of a site are provided there is still a risk of search engines that do index frames directing users to an incomplete frame page instead of a more appropriate equivalent non-frames page. If the frames pages of a site are ignored then this may mean that information deeper inside will not be available for users to search for, potentially a problem. For some sites this will not be an issue and they can have a legitimate claim to use frames. Examples are sites that have content changing too frequently to be picked up by search engines and sites that have restricted access of one kind or another. Sites can also maintain a duplicate, equally high quality non-frames version for this purpose, or use a non-frames front page to court search engines.

The figures for use of the more recent and advanced programming languages Java and JavaScript on site front pages show an interesting spread across the world. JavaScript is more widely used, often for graphics-enhancing effects such as the image rollover where an image on a page changes in response to the mouse passing over it. Java is used for a range of tasks, from graphics special effects to navigation buttons and continuous news tickertapes. Both of these languages do not operate in older browsers and users with newer browsers may choose to switch them off for security reasons and so they should not usually be used for essential site functionality. For this reason their inclusion in a site is not necessarily a good thing. A comparison of the rates of use for different countries shows a spread of results and it is certainly not the case that smaller domains or domains where the Internet is relatively underdeveloped are using them less.

Links between the presence of the five design features and registration in search engines were investigated by cross-tabulating the whole data set. A test of significance was not calculated because the data was not selected randomly. There was however an apparently strong link between the use of keywords and description META tags and registration in the search engines. This does not prove that these tags

help register the page, an alternative explanation is that those including the tags are more likely to also take the other steps necessary to get their site registered. In some countries, such as France and the UK, there is good use of META tags, but poorer coverage by search engines. This may reflect an extra complicating factor such as a popular national search engine, such as www.yell.co.uk in the UK or www.hachette.net in France, which may be the focus of national business coverage. Java and JavaScript also showed a positive link to search engine registration, although a much smaller difference was evident. A similar explanation is also likely here: designers who used extra features in their pages may have taken more trouble over registering them. A more surprising result was the small correlation between using frames and all search engines except InfoSeek, for which there was a negative correlation. It would be expected however that there would be a strong negative link between the use of frames and deeper indexing of the web site in some of the search engines. Unless there is an unknown common factor, it seems that InfoSeek may not favour frames based sites in its registration and retention algorithms.

The individual variations between countries for the design features columns of the table may be explained by a range of factors including the biases in the survey and 'house styles' of large or influential web designers in different countries. They will also reflect to an extent any nationally dominant web page design software, particularly in non-English speaking countries where there may be only a limited set made available in or translated into the local language.

Conclusions

The survey has produced results that should be disturbing reading for web site owners internationally, with nearly a quarter of sites surveyed not registered in any of the five major search engines that were tested. It is believed that this shortfall is probably caused in the main by misunderstanding of the importance of search engines by web site designers or by a lack of knowledge of how to get sites registered and to remain registered. Although the issue of search engine registration is not straightforward because the size of the web forces search engines to be selective about the pages that they choose and the algorithms that make the selection are not fully published, the widespread lack of META tags for contents and description, in addition to the use of frames based home pages are clear indications that search engines are not being taken seriously. Although the main search engines are not the only way to bring users into a site, they are free to register in and have a large user base, and therefore for most sites there is no reason to ignore them. The fact that 77% of sites are registered in at least one search engine means that the majority of sites should be able to be found using meta-search engines, which compile the results of a number of other search engines to give greater overall coverage. This should, however, not give much comfort to the site owners because these do not as yet have a large proportion of search engine users, all being outside the top ten search engine lists (CyberAtlas, 1999).

The search engines tested showed clear national coverage differences, greater than would be expected from their differing speeds and size alone, with evidence of probably unintentional domain bias in their page finding or page retention algorithms. It is stressed that this, and the rest of the paper, is in no way an assessment of the quality of the search engines themselves, the focus is on the sites surveyed and not the needs of the search engine user.

The differences in the results across the domains surveyed show a varying rate of search engine registration success. The com domain scores well in all relevant categories, as is to be expected from it as the sought after default domain for many

browsers. The various national domains show a large variation in most categories, much more than could be attributed to chance. Their different uses of web page design features may reflect national trends or nationally successful software products to an extent, but the results for search engines clearly show the less developed domains performing poorly. This may reflect either bad practice of individuals, a lack of development of national web infrastructure in the form of link sites or official gateway sites, a lack of 'sophistication' in ISPs and design companies in not providing automatic registration services, or for small domains the lack of a necessity to relate to search engines because other methods of publicising the address are sufficient. Countries that score well with search engines may also do so for incidental reasons, such as having linked multiple different language versions of many sites, or for more systematic reasons such as having well organised gateway sites. Countries may also score badly because the international search engines are not as important as elsewhere, either because of well-ordered gateway sites, or because of the existence of popular national search engines.

There is clearly much room for improving the accessibility of commercial web sites. It is believed that there is a strong case for national initiatives in creating gateway sites and in promoting good practice in relation to designing sites with web search engine retrieval in mind. Countries that do this have the potential to secure a head start at this crucial phase in the commercial development of the Internet.

References

- 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.
- Chun, T. Y. (1999), "World Wide Web Robots: An Overview", *Online & CD-ROM Review*, Vol. 23 No. 3, pp. 135-142.
- CyberAtlas (1999), <http://www.cyberatlas.com>, (accessed 10 July 1999).
- Dowe, D.L. Allison, L. and Pringle, G. (1998), "The Hunter and the Hunted - Modelling the Relationship Between Web Pages and Search Engines", *Lecture Notes in Artificial Intelligence* 1394, pp. 380-382.
- Gordon, M. and Patak, P. (1999), "Finding Information on the World Wide Web: the retrieval effectiveness of Search Engines", *Information Processing and Management*, Vol.35, pp. 141-180.
- Henzinger, M.R., Heydon, A., Mitzenmacher M. and Najork, M. (1999), "Measuring Index Quality using random walks on the Web", *Computer Networks and ISDN Systems*, Vol. 31 No. 11-16, pp. 1291-1303.
- Kirsch, S. (1998), "Infoseek's experiences searching the Internet", *SIGIR Forum*, Vol. 32 No. 2, pp. 3-7.
- Laursen, J.V. (1998), "Search Engine Persuasion", *Database*, Vol. 21 No. 1, pp. 42-46.
- Lawrence, S. and Giles, C. L. (1998), "Searching the World Wide Web", *Science*, Vol. 280, pp. 98-100.
- Lawrence, S. and Giles, C. L. (1999), "Accessibility of information on the web", *Nature*, Vol. 400, pp. 107-109.
- Marchiori, M. "Security of World Wide Web Search Engines", in: D. Gritzalis (Ed), *Reliability and Safety of Software Systems*, Chapman and Hall, 1997, pp. 161-174.

- Middleton, I., McConnell, M. and Davidson, G. (1999), "Presenting a model for the structure and content of a university World Wide Web site", *Journal of Information Science*, Vol.25 no.3, pp. 219-27.
- Northern Webs, Search Engine Tutorial for Web Designers, <http://www.northernwebs.com/set>, accessed 5 August 1999.
- Pardun, C.J. and Lamb, L. (1999), "Corporate Web sites in traditional advertisements", *Internet Research*, Vol. 9. No. 2.
- Pringle, G., Allison, L. and Dowe, D. L. (1998), "What is a tall poppy among web pages?", *Computer Networks and ISDN Systems*, Vol. 30 No. 1-7, pp. 369-377.
- Schwartz, C. (1998), "Web Search Engines", *Journal of the American Society for Information Science*, Vol. 49 No. 11, pp. 973-982.
- Simeon, R. (1999), "Evaluating domestic and international Web-site strategies", *Internet Research*, Vol. 9 No. 4, pp. 297-308.
- Snyder, H. and Rosenbaum, H. (1998), "How Public is the Web?: Robots, Access and Scholarly Communication", *Proceedings of the ASIS 98 Annual Meeting*, pp. 453-462.
- Spink, H., Bateman J. and Jansen, B.J. (1999), "Searching the Web: A survey of EXCITE users", *Internet Research*, Vol. 9 No. 2, pp. 117-128.
- Thelwall. M., (1999), "Business use of the .co.uk domain", University of Wolverhampton.
- Tunender H. and Ervin J. (1998), "How to Succeed in Promoting Your Web Site: The Impact of Search Engine Registration on Retrieval of a World Wide Web Site", *Information Technology and Libraries*, Vol. 17 No. 3, pp. 173-179.

Table I Search engine coverage

Country	Domain	Ending	Real Sites	Yahoo	Hotbot	AltaVista	MSN	InfoSeek	All	None
Finland		fi	562	70%	66%	82%	82%	49%	28%	7%
World	com	com	8050	62%	59%	81%	71%	67%	37%	9%
Switzerland		ch	2932	66%	61%	79%	62%	32%	17%	11%
Sweden		se	1827	63%	59%	75%	64%	48%	24%	13%
Japan	co	co.jp	2497	17%	16%	74%	30%	33%	6%	17%
Singapore	com	com.sg	380	63%	58%	69%	51%	35%	16%	18%
Russia		ru	1072	32%	30%	68%	37%	33%	9%	19%
France		fr	1483	66%	61%	68%	61%	34%	18%	16%
Holland		nl	3392	58%	54%	67%	51%	36%	18%	17%
Ireland		ie	438	53%	49%	67%	56%	49%	26%	22%
Belgium		be	987	52%	49%	67%	50%	34%	17%	23%
Israel	co	co.il	438	52%	50%	66%	60%	42%	18%	18%
Czech Republic		cz	1456	41%	37%	66%	47%	17%	7%	23%
Spain		es	980	54%	48%	66%	54%	42%	23%	24%
Portugal		pt	455	55%	50%	64%	54%	35%	20%	25%
Italy		it	2166	57%	53%	63%	49%	39%	19%	19%
Thailand	co	co.th	230	41%	37%	63%	42%	34%	14%	32%
Germany		de	7350	55%	52%	62%	52%	30%	16%	25%
Australia	com	com.au	2849	50%	46%	61%	49%	73%	24%	12%
Turkey	com	com.tr	387	44%	41%	59%	41%	22%	8%	29%
Iceland		is	312	55%	52%	59%	54%	30%	17%	27%
Greece		gr	355	52%	47%	58%	48%	28%	17%	31%
Taiwan	com	com.tw	1182	50%	49%	58%	51%	25%	13%	29%
UK	co	co.uk	3643	45%	41%	57%	45%	49%	19%	27%
Norway		no	1192	49%	45%	57%	48%	27%	14%	31%
New Zealand	co	co.nz	641	47%	43%	56%	47%	36%	16%	31%
Denmark		dk	2762	52%	48%	56%	49%	24%	12%	29%
Poland	com	com.pl	975	55%	50%	51%	46%	20%	9%	24%
Brazil	com	com.br	2584	36%	33%	51%	29%	20%	8%	39%
South Africa	co	co.za	1270	38%	36%	49%	37%	31%	13%	40%
Austria	co	co.at	369	43%	38%	48%	46%	30%	12%	37%
Peru	com	com.pe	51	39%	33%	47%	39%	25%	12%	37%
Philippines	com	com.ph	92	32%	29%	46%	33%	22%	12%	43%
Argentina	com	com.ar	632	34%	32%	46%	31%	26%	12%	45%
Indonesia	co	co.id	144	66%	53%	45%	59%	28%	12%	24%
Pakistan	com	com.pk	108	30%	26%	44%	26%	26%	11%	46%
South Korea	co	co.kr	1990	37%	33%	44%	32%	15%	7%	45%
Malaysia	com	com.my	327	35%	30%	43%	30%	25%	9%	43%
India	co	co.in	29	24%	21%	41%	21%	21%	10%	55%
Mexico	com	com.mx	624	34%	32%	40%	30%	25%	11%	47%
China	com	com.cn	815	35%	33%	39%	34%	10%	5%	47%
Egypt	com	com.eg	59	29%	29%	37%	25%	22%	10%	54%
Average				51%	47%	64%	51%	38%	18%	23%

Table II Selected HTML features found in sites checked

Country	Domain	Ending	Sites	Real Sites	% Key- Real words	Descr- ption	Frames	Java- Script	Java	
France		fr	1824	1483	81%	53%	50%	34%	24%	6%
Ireland		ie	519	438	84%	50%	47%	24%	31%	10%
UK	co	co.uk	5138	3643	71%	48%	45%	24%	28%	8%
Israel	co	co.il	520	438	84%	47%	45%	20%	32%	13%
Germany		de	8553	7350	86%	47%	43%	35%	32%	7%
Sweden		se	1958	1827	93%	45%	42%	37%	28%	5%
World	com	com	14884	8050	54%	44%	40%	17%	30%	8%
Switzerland		ch	3188	2932	92%	42%	38%	29%	27%	8%
Australia	com	com.au	3457	2849	82%	40%	38%	22%	29%	7%
New Zealand	co	co.nz	886	641	72%	40%	36%	19%	27%	7%
Austria	co	co.at	477	369	77%	37%	33%	30%	31%	6%
Holland		nl	3925	3392	86%	36%	40%	36%	35%	8%
Belgium		be	1111	987	89%	36%	31%	24%	29%	7%
Poland	com	com.pl	1187	975	82%	35%	29%	23%	28%	5%
Russia		ru	1195	1072	90%	34%	29%	17%	36%	6%
India	co	co.in	43	29	67%	34%	28%	31%	48%	7%
South Africa	co	co.za	1692	1270	75%	33%	30%	28%	27%	10%
Finland		fi	617	562	91%	32%	25%	27%	23%	3%
Singapore	com	com.sg	527	380	72%	30%	31%	21%	33%	11%
Malaysia	com	com.my	439	327	74%	30%	30%	18%	36%	10%
Pakistan	com	com.pk	122	108	89%	30%	24%	8%	25%	18%
Greece		gr	405	355	88%	29%	28%	23%	28%	8%
Turkey	com	com.tr	480	387	81%	29%	27%	20%	25%	10%
Philippines	com	com.ph	114	92	81%	28%	29%	18%	33%	12%
Spain		es	1090	980	90%	28%	27%	24%	31%	8%
Norway		no	1376	1192	87%	29%	26%	37%	23%	7%
Denmark		dk	3470	2762	80%	28%	26%	35%	25%	9%
Thailand	co	co.th	278	230	83%	27%	25%	24%	33%	10%
Argentina	com	com.ar	784	632	81%	26%	25%	23%	28%	12%
South Korea	co	co.kr	2262	1990	88%	23%	26%	27%	39%	6%
Mexico	com	com.mx	753	624	83%	22%	22%	18%	32%	10%
Peru	com	com.pe	58	51	88%	22%	22%	25%	41%	6%
Egypt	com	com.eg	75	59	79%	22%	19%	14%	20%	14%
Italy		it	2293	2166	94%	20%	29%	20%	25%	9%
Indonesia	co	co.id	185	144	78%	19%	19%	23%	21%	7%
Czech Republic		cz	1700	1456	86%	19%	15%	23%	27%	5%
Iceland		is	361	312	86%	19%	16%	29%	24%	5%
Brazil	com	com.br	3064	2584	84%	15%	14%	30%	30%	12%
Japan	co	co.jp	3007	2497	83%	15%	13%	22%	25%	3%
Portugal		pt	551	455	83%	15%	13%	26%	30%	13%
Taiwan	com	com.tw	1306	1182	91%	13%	13%	24%	29%	12%
China	com	com.cn	938	815	87%	8%	7%	19%	29%	11%
Average						35%	33%	26%	29%	8%