European Union Associated University Websites[1]

Mike Thelwall, Ray Binns, Gareth Harries, Theresa Page-Kennedy, Liz Price and David Wilkinson.
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1SB, UK.

Abstract

The web site is an important communication medium for universities in many countries. There are numerous reasons to expect that their characteristics will vary along national lines, the most immediate being differences in technological level and the organisation of higher education. In a world where the web is seen in many places as an important source of information it has the potential to overcome national boundaries, but are there still technological barriers? This paper reports on the results of a survey of the sizes of 670 web sites of higher education institutions in countries associated with the European Union, as estimated by AltaVista. It finds that there are still enormous national differences of up to three orders of magnitude. A related issue addressed is the extent to which AltaVista's coverage of university web sites is reliable and consistent across Europe. Large but uneven differences were identified between the main engine and national variations. Despite such methodological problems and cultural reasons for national variations in web site development, a clear pattern emerges, with the richer countries in Europe having much larger web sites. This is a problem for those wishing to use the Internet to increase international collaboration.

Introduction

The web is now an integral part of academic life in many countries. At a university in the UK, for example, a student may find that many of their studies have an online component, and some may even be taught completely online. Those researching their subject may well also see the web as a natural source of additional information, and others will be required to create sets of web pages for assignments. Providing scholarly and pedagogical information is, however, only one of the functions of a university web site (*Middleton* et al., 1999). It is also a marketing tool for the university as a whole, giving information to prospective students about the courses available and the university itself. In addition, there is likely to be information about the research conducted at the university, including details of active scholars, groups and projects. If such information is indexed in a search engine then it provides a new route for the discovery of others' research activities. As an example of this, a professor at Wolverhampton University was approached by a publisher with an academic book proposal, having found him by a web search for his specialism. Since search engines and the web are to a significant degree international, they can be agents to facilitate increased choice for students outside their native country. The same is true for academics seeking research partners, for example those wishing to build collaboration between different disciplines and European Union (EU) associated countries in order to bid for EU research funding (*Europa*, 2001). The potential

[1] Thelwall, M. Binns, R. Harries, G. Page-Kennedy, T. Price E. and Wilkinson, D. (2002). European Union associated university websites, Scientometrics, 53(1), 95-111.

research benefits of internationalisation through the Internet have been recognised before the age of the web (*Gruntorad*, 1992) but it is, conversely, equally a threat to those countries that are not using it effectively. It is important to know whether significant use of the web for information retrieval will leave some countries behind and provide a barrier to their integration in Europe, at least at the level of higher education.

The set of countries associated with the EU has expanded in recent years to include many from Eastern Europe. From this area in 1998, Russia, Poland, the Czech Republic, Hungary, Estonia and Slovenia were believed to be regional Internet leaders (*Sroka*, 1998), and it would be interesting to see if this appeared to be evident in academia 3 years later. There do not appear to have been any other surveys of Europe-wide university web site sizes. There have, however, been many studies analysing individual university web sites. One issue that has arisen is that some universities have not allowed their sites to be indexed by search engines (*Snyder* and *Rosenbaum*, 1998), although this does not appear to be the case for any complete sites in the UK (*Thelwall*, 2001b). If this practice were found to be common in any given country then it would be a cause for concern, making that country's higher education less 'visible'. Another factor that may make sites less well indexed is the use of dynamically created HTML pages instead of static files. Such pages can be ignored by search engines as potential causes of spurious index entries (*Thelwall*, 2001a) and so this relatively high technology type of site could reduce a university's web profile.

A survey will be described of university web sites in EU-associated countries, the objective of which was to identify national trends and causes for concern from an information retrieval point of view. In addition to a consideration of the reliability of results, there will be three major issues addressed. The first is whether there are national differences in web site sizes. These are relatively simple statistics, yet may give an indication of the amount of information available. The second issue is one of visibility: how well the sites are covered by some major search engines. Finally, linguistic variations will be tackled, driven by the consideration that language can be a barrier to some (*Large* and *Moukdad*, 2000). The patterns identified will then be compared with national information external to the web in order to provide suggested explanations.

The Structure of Higher Education in Europe

Higher Education is far from uniform across Europe, indeed it has been said that, including private education, there are more systems than countries (*Knudsen* et al., 1999). There is a basic divide between countries like Germany and the Netherlands that have a binary structure and those like the UK and, to some extent, Sweden that have essentially a single level of organisation. In Germany, for example, the main distinction is between universities, which give a broad education and award higher degrees, and Fachhochschulen, which are non-university higher education institutions that give a professional and vocational education but do not award higher degrees. In the UK, there is essentially a single system of universities. The old apparently binary system of universities and the more vocationally oriented polytechnics was scrapped in 1992 with polytechnics becoming universities. It was, nevertheless, already different from the classic binary pattern because the old polytechnics delivered a broad range of education and conducted research, although there was still an overall vocational bias compared to the old university sector (*Wright* et al., 1997a). This clearly demonstrates the complexity of higher education, even in a single country. The types of national diversity include the degree of state involvement, the national uniformity of qualifications and the number and status of private universities. The diversity of access methods (*Boezerooy* and *Vossensteyn*, 1999) award duration and levels throughout Europe, with consequent misunderstandings of each other's institutions (*Wright* et al. 1997b) have led to attempts to introduce standardisation to facilitate the interchangeability of qualifications. At ministerial level this has led to such statements of intent as the *Bologna Declaration* (1999), although it has been claimed that a rigid system of qualifications is neither desirable nor feasible and that real variations are likely to persist (*Knudsen* et al., 1999). Educational institutions have also recently being going through a process of change, partly in response to increasing globalisation and the impetus of information technology (*Sporn*, 1999). Europe, then, is an area in which there are two major approaches to higher education but many variations of each, despite some movement towards standardisation. Any attempt to make comparisons across the region must, of necessity, allow compromises to accommodate the real differences that exist. It must similarly be the case that the findings of such studies will be weakened by the situation.

An additional problem with comparing sizes of institutions between countries is the extent of fragmentation of higher education. It is common for countries to have multi-faculty universities in combination with more specialist institutes, but in some this is more pronounced than in others. There is also a practical problem with identifying in each country a precise list of all institutions that have any given international standing. This is especially problematic as terms used have different meanings, for example a polytechnic is a university level institution in France, but in Finland it would be of lower status. A good source of authoritative information is the European Union report (1998), although this does not give a list of names of university status institutions in all countries.

Methodology

The first task was to identify the websites of all universities in EU-associated countries. This definition includes all countries that are members of the EU or were identified as eligible for some kind of special treatment for EU research funding in early 2001 (*UK Research Office*, 2001). In order to concentrate as far as possible on similar entities, consideration was to be restricted to multi-disciplinary universities, excluding any with a very narrow focus. Any exceptions to this rule will be noted, with reasons. The process of identifying qualifying institutions was started by consulting official (*European Union*, 1998) and unofficial (*EuroEducation.Net*, 2001) sources of information about the structure of Higher Education (HE) in each country. From these, the naming conventions to identify different levels of institution were identified. Various online international and national lists of higher educational institution home pages were then used to identify the actual universities and their web sites. These lists came from various sources, both official and unofficial: search engine university categories; individually maintained lists; and government information pages. It had been decided to focus upon the research aspect of education and, therefore, only to include institutions with a significant research profile. Such a selection criteria would produce a useful starting list for those wishing to seek research partners in other institutions in Europe. Indeed, at least one university has published an explicit requirement for 'high standing' as a requirement even for teaching collaboration (*Cambridge University*, 2001). An additional step was then necessary: to identify institutions meeting this criterion. It was decided to use the UK university status as the benchmark for comparison. This meant that the UK would include 98 universities but not non-university institutions, including colleges of further and higher education, although these can and do deliver higher education. In order for a UK HE college to become a university it must satisfy several criteria, including the production of research and the supervision of Ph.D. students. These were the key pieces of information to be used to decide which organisations to include from other countries. It was necessary to invest significant time to decide this issue because of the differing structures and naming systems for HE in Europe.

An initial selection procedure had to be employed to make the task of analysing the websites manageable, since there were at least 401 institutions in England alone receiving research funding (*HEFCE*, 2001). The approach adopted for university classification was to start with a list of all HE institutions, and then to reject those that were not called ‘universities’ or ‘institutes’. This rather arbitrary decision was necessary to reduce the numbers to a manageable level and was arrived at after an initial survey of the national naming conventions and selected sites. This rule, if applied to the UK, would have produced a slightly different list, for example including Bolton Institute of Higher Education, which does not currently have university status, although it has applied for it (*House of Commons*, 2000). Following this step, which left around 1,000 sites, the countries were examined individually for naming conventions in order to establish patterns to further identify university-equivalent institutions. The examination involved visiting selected institutional web sites for self-descriptive information, particularly about research and Ph.D. supervision, and attempting to find authoritative national education information sources.

Once the list of universities had been finalised, the next stage was to analyse their web site sizes. *Bar-Ilan* (2001) has suggested the creation of an information science search engine as a reliable data source for Informetrics and this would be an ideal solution for this type of exercise, but it is not available yet. The Internet Archive (www.archive.org) does offer access to a historical database of the web, and is a promising new resource, but its primary data source, at the time of the study was the database of one of the less well-known crawlers. It was decided to use commercial search engines that provide site count data. Search engine counts have been found to be unreliable (*Bar-Ilan*, 1999; *Rousseau*, 1999; *Snyder* and *Rosenbaum*, 1999; *Thelwall*, 2000c) although recent improvements have been identified in AltaVista, (*Thelwall*, 2001a; *Thelwall*, 2001b) which make it the most suitable. AltaVista, Go and HotBot were all used on the full data set but the counts from AltaVista were larger than those from the others and so the other figures were not able to provide any additional useful information and AltaVista was chosen as the sole source. Google appeared to be the largest search engine on the web but, although it will search individual sites upon request, it will not report a simple count of pages and, therefore, could not be used. The first question to be asked about the search engine data was how well it matched reality. To decide this, selected sites were visited to judge whether the results were correct. All sites reported with low page counts were visited, plus a selection of others. This process led to the discovery, as a by-product of the main purpose, that some domain names were obsolete or incorrect, and these were corrected. The manual check was also used to ascertain whether there were any identifiable linguistic, national or technological factors influencing the degree of site coverage by the search engines.

AltaVista provides many geographic variations of its search service, for example a French version that searches only web sites based in France. These may well give greater coverage of web sites within the national boundaries, but their use as a primary data collection tool when multiple countries are covered is problematical, since they are not available for all countries. One method that will be used to test the reliability of the main AltaVista is to compare variations between results for national and global search engines for the universities studied. The exact relationship between the two may shed some light on the extent of coverage of web sites by both, and AltaVista's policy for coverage of domains when it has a regional variant in place, presumably with prime responsibility for that domain. The conclusions will be of particular interest to those wishing to count web links between countries, for example for Web Impact Factor and other calculations (*Ingwersen*, 1998; *Leydesdorff* and *Curran*, 2000, *Darmoni* et al., 2000).

Finally, an analysis of patterns in web site sizes throughout Europe was conducted, focussing on identifying national variations.

Results and Discussion

Technical Issues in Site Coverage

There were a number of web sites that were large but only had a few pages indexed by AltaVista, including the University of Glamorgan, the University of Abertay, the University of Bacau, Siauliai University, Escuela Univeritaria and Università degli studi di Macerata. All of these used the HTML frameset feature, which splits each page into different files and is impossible to index accurately in a search engine. It is hypothesised that this is the reason for their low coverage. A visit to the Macerata web site by a personal crawler showed that it contained 1,686 distinct HTML pages, but the main AltaVista reported "about 80" although the Italian AltaVista achieved much better coverage, with "1289 pagine trovate". In this case, the national crawler was much more successful in covering the site. A Google search for the common Italian word "di" on this web site found "about 423" pages, and the relative commonness of this word suggests that the actual coverage by Google is approximately 550, indicating that it, too, is having difficulty in covering the site comprehensively. The personal crawler used for our study did not need to process and index the pages, making its task much simpler than that of a search engine. For Glamorgan, in contrast, the main engine returns 4 pages but the national none, although the site is clearly a large and complex one. Five other small sites were incompletely indexed for other reasons as described below.

- The International University of Social Sciences (LEX) in Estonia had 19 pages but only one indexed in AltaVista, for an unknown reason.
- University of Travna has random pages missing for an unknown reason.
- The Technical University of Petrosani had only one page indexed. It is hypothesised that this was due to an unusual referring convention in the HMTL of many official pages, which was the (unnecessary) use of a dot to start relative path names. This worked in the browsers in which it was tested, but AltaVista does not seem to recognise the targeted pages.
- The University of Targu Jiu has a count of zero despite having many pages. It is hypothesised that AltaVista has ignored it because the home page title is 'Test Page'.
- The European University of Lefke was sometimes reported as having 3 pages by AltaVista, and sometimes 15, even on the same day. The latter is a better estimate but the reason for the lower figure is unknown. Google managed at least 223

pages (in a search for “lefke +site:lefke.edu.tr”) and a visit by our crawler counted 251.

- The Danmarks Pædagogiske Universitet was reported as having no pages, which may be a result of its relative youth. It was perhaps no older than 7 months at the time of testing.

France posed a unique problem for the analysis as a result of individual university web sites being often derived from the group of universities of which they are a member. This resulted in long domain names and often departments having domain names not derived from the official university one, but from the university group. As an example of this, the Université Henri Poincaré (Nancy 1) in the Université de Nancy has domain name root uhp-nancy.fr, but some of the parts of its university web site are derived from the university group web site u-nancy.fr, rather than uhp-nancy.fr. Examples of these are: Centre de Médecine Préventive (www.cmp.u-nancy.fr); École de Santé Publique de Nancy (www.sante-pub.u-nancy.fr/esp/); Faculté des Sciences - Bibliothèque (www.sciences.bu.u-nancy.fr) (*StanNet*, 2001). Some shared facilities are also hosted on the university group web site. This means that the non-derivative pages will not be included in the Nancy I page count, which is based upon the root uhp-nancy.fr, and it will, therefore, be smaller as a result. The widespread appearance of similar phenomena in France is a likely contributory factor to its relatively low site counts. Other countries are not exempt from this problem, for example the university of Manchester (man.ac.uk, manchester.ac.uk) has a non-derived name for its computing centre (mcc.ac.uk) but this type of singularity appears to be much more widespread in France than elsewhere.

Linguistic Considerations

AltaVista was created in the United States of America, and so one cause for concern was whether it would be capable of indexing non-English sites, particularly those with a non-ASCII character set, such as is used in parts of Estonia, for example. The survey showed no evidence of language bias and proved its ability to index non-ASCII character sets such as Cyrillic for Russian language pages. There were also no apparent national biases for AltaVista, even in countries like Romania that seemed to have very slow Internet links, at least to the UK. Some countries, such as Switzerland and the Netherlands, had a tendency for the official part of their websites to be offered in multiple languages. Such page replication could be expected to have an impact on overall sizes. This may not be great, however, since the need for multilinguality may actually inhibit page creation and it may also not affect unofficial areas.

Page Counts Reported by AltaVista

Table 1 gives a summary of the sizes of the web sites of the universities identified in each EU associated country, as reported by the main AltaVista. It is ordered by median site size, the best measure of central tendency for this kind of data. The information for per capita Gross Domestic Product (GDP) and population are taken from the CIA World Factbook 2000 (*CIA*, 2000).

Median	Name	Sites	Smallest	Largest	Median	Mean	GDP/cap	Population	Code
rank							($)		
1	Norway	4	40209	151101	122976	109316	25100	4481	no
2	Germany	45	7585	278213	64304	79201	22700	82797	de
3	Netherlands	15	705	188226	49170	58692	23100	15892	nl
4	Sweden	14	6532	290887	34871	71027	20700	8873	se
5	Finland	10	7232	247675	31345	55103	21000	5167	fi
6	Belgium	9	836	61029	31248	28007	23900	10241	be
7	Austria	13	1426	172685	29732	56756	23400	8131	at
8	Slovenia	2	15005	39088	27047	27047	10900	1927	si
9	Denmark	5	6648	173850	23882	49908	23800	5336	dk
10	Britain	98	4	287224	19781	37105	21800	59511	uk
11	Ireland	7	4962	39293	17673	19737	20300	3797	ie
12	Switzerland	10	95	114413	15477	26159	27100	7262	ch
13	Israel	6	2680	171727	15151	44434	18300	5842	il
14	Portugal	12	291	71320	11123	16688	15300	10048	pt
15	Spain	59	0	47236	6230	10317	17300	39997	es
16	Italy	47	17	66863	5912	11087	21400	57632	it
17	Czech Rep.	16	0	204374	5737	22347	11700	10272	cz
18	Greece	17	11	24020	4408	6914	13900	10602	gr
19	Poland	16	104	18136	3772	5603	7200	38646	pl
20	Hungary	19	134	215424	3469	20392	7800	10139	hu
21	Cyprus	5	11	8175	3071	2990	15400	758	cy
22	France	72	0	52634	2988	6373	23300	59330	fr
23	Slovak Rep.	9	3	43844	1797	7488	8500	5408	sk
24	Lithuania	3	671	3254	1532	1819	4800	3601	lt
25	Estonia	8	0	177432	1068	24424	5600	1431	ee
26	Latvia	5	30	9980	957	2748	4200	2405	lv
27	Luxembourg	1	939	939	939	939	34200	437	lu
28	Iceland	3	168	3228	571	1322	23500	276	is
29	Liechtenstein	1	285	285	285	285	23000	32	li
30	Romania	20	0	13031	196	1075	3900	22411	ro
31	Bulgaria	17	0	5071	61	472	4300	7797	bg

Table 1. Web Site Sizes for Universities in European Union Associated Countries

The universities in all countries for which there was a national variant of AltaVista were rechecked through this alternative route. In Spain, Belgium and Switzerland there were different language interfaces available, but these all gave the same search results. Some national variants reported greater coverage than the main AltaVista, but others reported much less, as shown in table 2. It is possibly the case that older national portals have greater coverage through retention of URLs from previous crawls (*Thelwall*, 2001b), or simply as a result of having greater storage space.

Domain	National search engine median divided by global search engine median
at	0.2
be	0.2
ch	0.3
ie	0.3
no	0.3
pt	0.6
dk	1.0
es	1.0
fr	1.1
se	1.8
it	3.9
de	5.7
uk	6.5
nl	17.5

Table 2. A comparison of median reported web site sizes for universities in EU Associated Countries for which AltaVista provides a national variant.

Table 3 is based upon the largest overall median results, irrespective of source. This is unsatisfactory from the point of view of using a non-uniform source of information. Assuming that web crawlers cover a fraction of a site and report their results reasonably accurately, all reported figures would all be underestimates of the actual page counts and so these new figures would be individually more reliable. In support of this hypothesis, the Dutch university with the largest page count was examined in more detail and it was found that the main AltaVista had ignored most of some sections of the site. For example, its reported count of www.kbs.twi.tudelft.nl, a huge sub-site that contained many mirrors of computer documentation, was clearly an underestimate. It may be the case that Dutch universities do contain many mirrors of computer documentation that the main AltaVista ignores, either because it has indexed it elsewhere, or as part of a general policy of not indexing too many pages on a given sites, or for another unknown reason. This would support the hypothesis that the national counts were more accurate. Table 1 will, nevertheless, be the one used for further analysis in order to provide a consistent data source. It's suspected inaccuracy is noted, however.

Median rank	Name	Sites	Smallest	Largest	Median	Mean	National used?
1	Netherlands	15	1017	10799035	859927	2062055	Y
2	Germany	45	8006	1185882	364559	349723	Y
3	Britain	98	0	1378243	145011	209871	Y
4	Norway	4	40209	151101	122976	109316	N
5	Sweden	14	6061	262009	61480	96760	Y
6	Finland	10	7232	247675	31345	55103	
7	Belgium	9	836	61029	31248	28007	N
8	Austria	13	1426	172685	29732	56756	N
9	Slovenia	2	15005	39088	27047	27047	
10	Denmark	5	6648	173850	23882	49908	N
11	Italy	47	311	134186	22921	28978	Y
12	Ireland	7	4962	39293	17673	19737	N
13	Switzerland	10	95	114413	15477	26159	N
14	Israel	6	2680	171727	15151	44434	
15	Portugal	12	291	71320	11123	16688	N
16	Spain	59	0	47236	6230	10317	N
17	Czech Rep.	16	0	204374	5737	22347	
18	Greece	17	11	24020	4408	6914	
19	Poland	16	104	18136	3772	5603	
20	Hungary	19	134	215424	3469	20392	
21	France	72	0	409718	3270	27295	Y
22	Cyprus	5	11	8175	3071	2990	
23	Slovak Rep.	9	3	43844	1797	7488	
24	Lithuania	3	671	3254	1532	1819	
25	Estonia	8	0	177432	1068	24424	
26	Latvia	5	30	9980	957	2748	
27	Luxembourg	1	939	939	939	939	
28	Iceland	3	168	3228	571	1322	
29	Liechtenstein	1	285	285	285	285	
30	Romania	20	0	13031	196	1075	
31	Bulgaria	17	0	5071	61	472	

Table 3. Web Site Sizes for Universities in European Union Associated Countries - using the results of AltaVista or a national variant, whichever had the greater median.

Patterns in Site Size

The number of universities in each country clearly varies for a number of different reasons, one of these being the population of the country. Population was found to be highly correlated with the number of universities (Pearson: 0.853, significant at the p = 0.01 level). This is reassuring because it seems clear that larger universities would need more pages to describe their activities and, if the output per member of staff was the same, they would have proportionally more web pages than smaller universities. If it had found that the correlation between population and the number of universities was poor then it would be difficult to compare the actual sizes of their websites. Liechtenstein, Iceland and Luxembourg have very small populations in comparison with the remaining countries. Liechtenstein and Luxembourg have only one university each but the proximity of other European countries means that they are close to universities of other countries. Iceland, however, with a small population has three universities.

In general the median page count size was larger for the western half of Europe, with the exception of France, Liechtenstein, Luxembourg, Iceland and Slovenia, the latter having the largest page count for an Eastern European country.

Slovenia was the only Eastern European country in the top half of the table. However it has only two universities, both of which are large. Its GDP is high compared to other Eastern European countries. It is expected that countries with a higher per capita GDP would have more sophisticated web sites. GDP per capita against median page count size was correlated (Pearson: 0.407, significant at the p = 0.05 level; Spearman: 0.068, significant at the p=0.01 level). Ignoring the tiny countries of Liechtenstein, Iceland and Luxembourg, the correlation coefficients are 0.570 (Pearson, significant at the p = 0.01 level), and 0.406 (Spearman, significant at the p = 0.05 level). Similar calculations were carried out for the data from table 3 for comparison purposes. For this case the, parametric Pearson calculation did not find significant correlations, but the Spearman calculation did. The reason for this is the enormous value for Holland, which upsets a parametric calculation more than a non-parametric one because the latter is based only upon ranks. Figure 1 shows median page counts plotted against per capita GDP. From this it can be seen that Luxembourg is an outlier, but that a linear trend is present.

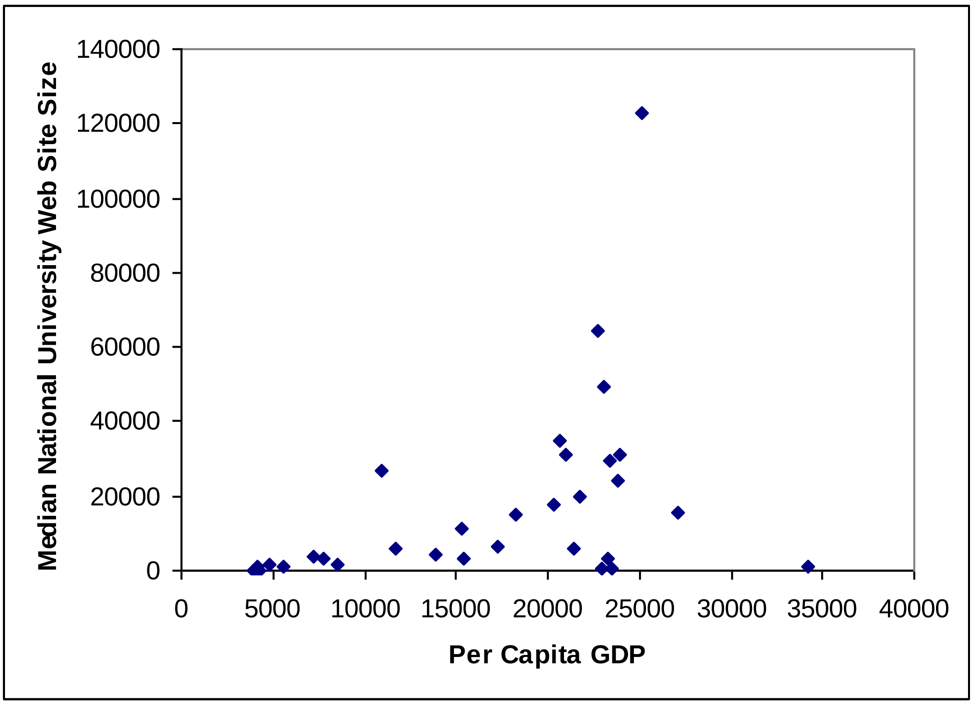

Figure 1: Median web site size against per capita GDP for European Union university web sites

The median page count for France was much lower than would be expected for its per capita GDP and was most remarkable in this respect, except for the smallest three countries. One factor contributing to the French site count 'shortfall' could be technical difficulties in identifying all sites associated with a French university, as described above. France and Hungary both had some very large web sites, although the majority were relatively small. There were considerable variations in the extent of coverage between the national and main AltaVista results at the individual university level. The count for the domain uhp-nancy.fr from the main AltaVista was, "About 2,121", for example, whereas the French AltaVista reported a much more respectable "135597 pages trouvées". The main reason for low coverage for this individual institution, then, appeared to be incomplete crawling by AltaVista. The reason for France's low median for web site size was the large number of web sites that returned very low counts in AltaVista. A sample of the small French web sites were examined in an attempt to ascertain the reason for low site counts.

- Montesquieu University - Bordeaux IV. The main AltaVista reports “About 436” pages, French AltaVista also reports 436. Our crawl reported 1,521 distinct HTML pages. The site uses frames. The university reports having 13,500 students, and so is not small. The university has some non-derivative domain names for affiliated web sites, including two institutes and two research centres (www.montesquieu.u-bordeaux.fr/dirweb.html).
- Paul Valery University - Montpellier III. The main AltaVista reports “About 560” pages, French AltaVista reports 561, and our crawler 11,133. The site uses advanced HTML, including JavaScript driven links, but not frames. The university reports having 20,000 students. The main site appeared to be very controlled, without links to gain access to teaching material or any other pages of individual members of staff, although AltaVista had found and indexed some of these pages. Information about other associated web sites could not be found. AltaVista had not found any pages on the maths server, math.univ-montp3.fr, and only two on metice.univ-montp3.fr. The latter site contained complex JavaScript filled HMTL, but even normal HTML links from indexed pages were not covered.
- Louis Pasteur University (ULP). The main AltaVista reports “About 865” pages, the French 1432 and our crawler 1642. The universities of Strasbourg have much information on a single server as a common point of access to all. There were several non-derivative domain names of institutions. No individual pages of members of staff were found in the results pages of AltaVista, all of the pages seemed to be official information pages.
- L'Université de la Méditerranée - Aix Marseille II. The main AltaVista reports “About 325” pages, French AltaVista reports 356 and our crawler 650. The main site contained general information about the departments (www.mediterranee.univ-mrs.fr/composantes/), but the departmental web sites had domain names that were derivative from the university group, for example www.ejcm.univ-mrs.fr, www.iut.univ-aix.fr, com.univ-mrs.fr.
- The University Jean Moulin - Lyon 3. The main AltaVista reports “About 537” pages, French AltaVista reports 998, and our crawler 4 (from the home page). The site was complex with frames and Java. Most pages contained official information although some contained conference information, and pages created by members of staff, apparently in an individual capacity (e.g. www.univ-lyon3.fr/siehldaweb/trevoux/ed-trevoux.htm).

Overall, French web sites seemed to suffer from two problems from an indexing point of view: the widespread use of non-derivative domain names, and frequent use of complex HTML. An impression was gained from the frequent lack of unofficial pages that, despite these compounding factors, there probably was a relatively low level of web page creation by individual academic staff members in French universities compared to, for example, the UK. This hints at a different cultural attitude to the web, or web page creation. A possible explanation for this is the domination of the web by English pages, and the cultural importance of French in France. This is perhaps well illustrated by the existence of a well-used word in French, ‘francophonie’, which translates as “the French speaking world” (Google reports about 159,000 pages containing this word). Simple counts from the main AltaVista did confirm a pattern of the relatively small size of the web in France. For example, it recorded 7,013,970 web pages of all kinds in France, 17,336,516 in the UK, and 31,152,229 in Germany.

Four countries stand out for having median page counts above the trend: Slovenia; the Netherlands; Germany; and Norway. The median of Slovenia, with only two universities, is unreliable, and will not be the subject of further comment. It is interesting that there is a reasonably consistent linear trend for per capita GDPs up to 20,000, whereas there is an enormous variation in median sizes between countries with very similar economic wealth. A possible explanation is that cultural factors predominate, given a rich enough economic base. The economic factor is likely to be a combination of the ability to afford the computing equipment and support staff, and the historical ability to have been able to afford it also in previous years, when it was more expensive, therefore developing expertise and a user base. If the latter is the dominant reason then much greater use of the web would be expected in the slightly less rich countries in future years. The larger points on the graph also suggest the presence of exponential growth in web use. Following the discussion of the cultural factors for France, it is possible that what is being illustrated is the exponential growth in *potential* web use, driven by economic (and historical economic) wealth.

Summary and Conclusions

The exercise of identifying university level institutions in EU-associated countries and interpreting the results from AltaVista was a complex one. The differing national higher education structures created problems in deciding upon the appropriate level of institution to include in each country. Once the final list of web site domain names had been identified, the coverage by the main version of AltaVista was checked on many of the smaller sites, and found to be demonstrably incomplete in a number of cases, mainly due to the complex HTML used in site design. A comparison of results between any nation variants and the main engine revealed large differences. These were uneven, with some national engines giving much greater median coverage but others much less.

In reported sizes of university web sites, there was a clear tendency for the richer countries to have larger university web sites. France was one of the exceptions to this, possibly due to its two-tier university structure, complications in domain name choices, and less creation of unofficial web pages. Some of the richer countries exhibited very high page counts, suggesting a possibility of exponential growth due either to economic factors directly or indirectly through a longer experience with the technology. The wide variation in median sizes for the richer countries was suggestive of cultural factors dominating the actual use of the web, rather than purely economic considerations.

In terms of the reliability of the results, it is evident that variations of an order of magnitude are possible between different versions of AltaVista, even for median university web site sizes. Further investigation of a Dutch web site indicated that the main AltaVista was not including all of its pages in the count. This issue merits further investigation for cybermetricians wishing to draw more detailed reliable conclusions about differences in site sizes between countries. It is particularly difficult to check independently, however, because of the ability of established crawlers to have greater coverage of a web site due to having a historical database of valid URLs, some of which may no longer be linked to, or not in a way that a spider can identify.

The difference in median site size between the largest and the smallest is enormous at 2,016 to 1, which must give a real cause for concern about the ability of the web to help cross the international divide in Europe. If the web does become the dominant medium for identifying the work of others then there is clearly a long way to go before this process is universally available throughout Europe, or even in EU-associated countries. Technology is developing in a way that should make cross-border collaboration easier, for instance with the existence of free online translation services, some associated with search engines. But the Internet as a source of inequality is far from being a new idea, but it is still seen as a real one for those seeking to promote international integration (*Chon*, 2001). To give a concrete example, scholars in Bulgaria are threatened with being collectively invisible on the web, and would have a case for EU assistance to avoid marginalisation.

References

BAR-ILAN, J. (1999). Search Engine Results over Time - A Case Study on Search Engine Stability. *Cybermetrics*, 2/3. [online]. Available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

BAR-ILAN, J. (2001). Data collection methods on the Web for informetric purposes - A review and analysis. *Scientometrics*, 50(1): 7-32.

BOEZEROOY, P., VOSSENSTEYN, H. (1999). How to get in? - a comparative overview of access to higher education. *Higher Education in Europe*, 24(3): 349-358.

BOLOGNA DECLARATION (1999). Joint declaration of the European Ministers of Education Convened in Bologna on the 19th of June 1999. [online]. Available: http://www.unige.ch/cre/activities/Bologna%20Forum/Bologne1999/bologna%20declaration.htm (March 2001)

CAMBRIDGE UNIVERSITY (2001). General Guidelines for consideration of proposals for collaboration with other universities in the provision of courses and examinations. [online]. Available: http://www.admin.cam.ac.uk/offices/education/collab.html (February 2001)

CHON, K. (2001). The future of the Internet digital divide, *Communications of the ACM*, 44(3): 116-117.

CIA (2000). World Factbook 2000. [online]. Available: http://www.odci.gov/cia/publications/factbook/index.html (May 2001)

DARMONI, S. J.. THIRION, B., DOUYERE, M., CHALLOUB, C., LEROY, J. P. (2000). Measurment of the Web site impact: the Web Impact Factor. The example of the French University Hospitals. Revue du Praticien - Médecine Générale 14 (516): 2079-2080.

EUROEDUCATION NET (2001). [online]. Available http://www.euroeducation.net/ (March 2001)

EUROPA (2001). Fifth framework programme. [online]. Available: http://europa.eu.int/comm/research/fp5.html (February 2001)

EUROPEAN UNION (1998). A guide to higher education systems and qualifications in the EU and EEA countries. [online]. Available: http://europa.eu.int/comm/education/socrates/erasmus/guide/default.html (March 2001)

GRUNTORAD, J. (1992). Research and academic networking in the Czech and Slovak Federal Republic, Computer Networks and ISDN Systems. 25(4-5): 438-443.

HEFCE, (2001). Institutions funded by the Council. [online]. Available: http://www.hefce.ac.uk/UniColl/default.htm (March 2001)

HOUSE OF COMMONS (2000). Commons Committee stage of the Learning and Skills Bill debate. [online]. Available: http://www.brad.ac.uk/admin/conted/action/News%20and%20Events/learningbill.htm

INGWERSEN, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54(2): 236-243.

KNUDSEN, I., HAUG, G., KIRSTEIN, J. (1999). Trends in Learning Structures in Higher Education. [online]. Available: http://www.rks.dk/trends1.htm (March 2001)

LEYDESDORFF, L., CURRAN, M. (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy. *Cybermetrics,* 4. [online]. Available: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html

MIDDLETON, I., MCCONNELL, M., DAVIDSON, G. (1999). Presenting a model for the structure and content of a university World Wide Web site. *Journal of Information Science,* 25(3): 219-227.

ROUSSEAU, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics,* 2/3. [online]. Available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

SNYDER, H., ROSENBAUM, H. (1998). How Public is the Web?: Robots, Access and Scholarly Communication. *Proceedings of the ASIS 98 Annual Meeting,* 453-462.

SNYDER, H., ROSENBAUM, H. (1999). Can search engines be used for web-link analysis? A critical review. *Journal of Documentation*, 55(4), 375-384.

SPORN, B. (1999). Towards more adaptive universities: trends of institutional reform in Europe. *Higher Education in Europe,* 24(1): 23-33.

SROKA, M. (1998). Commercial development of the Internet and WWW in Eastern Europe. *Online & CD-ROM Review,* 22(6): 367-376.

STANNET (2001), Les serveurs W3 sur StanNet. [online]. http://www.u-nancy.fr/ (April 2001)

THELWALL, M. (2001a). Results from a Web Impact Factor crawler. *Journal of Documentation,* 57(2): 177-191.

THELWALL, M. (2001b, to appear). Extracting Macroscopic Information from Web Links. *Journal of the American Society for Information Science and Technology*.

THELWALL, M. (2001c). The responsiveness of search engine indexes. *Cybermetrics,* 5(1). [online]. Available: http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html

UK RESEARCH OFFICE (2001). [online]. Available: http://www.ukro.ac.uk/

WRIGHT, P., CAMPBELL C., GARRETT, R. (1997a). Setting the context of higher education in Europe (The national committee of enquiry into Higher Education). [online]. Available: http://www.leeds.ac.uk/educol/ncihe/r11_065.htm (March 2001)

WRIGHT, P., CAMPBELL C., GARRETT, R. (1997b). The relationship of UK degrees to those elsewhere in Europe (The national committee of enquiry into Higher Education). [online]. Available: http://www.leeds.ac.uk/educol/ncihe/r11_069.htm (March 2001)