Libri, 2003, vol. 53, pp. 242–253
Printed in Germany · All rights reserved

Libri
ISSN 0024-2667

Disciplinary Differences in Academic Web Presence – A Statistical Study of the UK

MIKE THELWALL AND LIZ PRICE
School of Computing and Information Technology, University of Wolverhampton, Wolverhampton, UK

The Web has become an important tool for scholars to publicise their activities and disseminate their findings. In the information age, those who do not use it risk being bypassed. In this paper we introduce a statistical technique to assess the extent to which the broad spectrum of research areas are visible online in UK universities. Five broad subject categories are used for research, and inlink counts are used as indicators of online visibility or impact. The approach is designed to give more complete subject coverage than previous studies and to avoid the conceptual difficulties of a page classification approach, although one is used for triangulation. The results suggest that Science and Engineering dominate university Web presences, but with Humanities and Arts also achieving a high presence relative to its size, showing that high Web impact does not have to be restricted to the sciences. Research funding bodies should now consider whether action needs to be taken to ensure that opportunities are not being missed in the lower Web impact areas.

Introduction

The Web is becoming hegemonic as an interface for information, particularly that of an academic nature. Scholarly journals are increasingly available online, either in subscription-based publishers' digital libraries or posted in publicly accessible Web sites (Kling and Callahan 2004). Individual articles can also be found scattered around the Web on authors' Web sites or organised institutional repositories (Brody, Carr & Harnad 2002). Preprint archives, holding early versions of papers are a uniquely electronic communication medium and in some areas of science threaten to become hegemonic (Rees 2002). But, in addition to formal scholarly publications, the Web also hosts a wide variety of other academic information. Those searching a university site from a developed country are likely to find much research-related information such as online CVs for faculty, publication lists, and general research descriptions (Wilkinson, Harries, Thelwall and Price 2003) although the site may also host administrative, teaching and general information (Middleton, McConnell and Davidson 1999; Ferdig and Hartshorne 2002; Kebede 2002; McAvinia and Oliver 2002; Wang, Berry and Yang 2003).

On a smaller scale, any given research group may have its own mini-site describing its research, listing membership and publications, and perhaps also advertising for PhD students. Such online publicity must surely have some effect in raising the group's research profile, although it is not known how much effect or of what kinds. Nevertheless, Web use varies by discipline and those not using it should be concerned that they are not missing an opportunity. Self-publicity is an important multi-faceted activity (Hyland 2003) that is evolving in the electronic age. The successful researchers of the future may be those

Mike Thelwall is Head, Statistical Cybermetrics Research Group, School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. Tel: +44 1902 321470 Fax: +44 1902 321478. E-mail: m.thelwall@wlv.ac.uk
Liz Price. School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. Tel: +44 1902 321859 Fax: +44 1902 321478. E-mail: Liz.Price@wlv.ac.uk

who have the skills to harness it effectively. However, Web use is not well understood, both in terms of patterns of use and the implications of information availability to publishers and consumers. It would be valuable to know how useful and widespread Web use is for different academic subjects, particularly for scholarly communication outside of the host institution. Computer scientists have traditionally been heavy Web users, but have other disciplines now caught up or are they being left behind? An answer to this is essential for policy makers and research managers to assess whether the opportunities made available by the Web are fully exploited. In this study we use a new technique to address the question of whether all areas of scholarship have a significant Web presence, using the UK as a case study.

Literature review

Disciplinary differences in Web use

Early sociologists of the Web tended to make grand claims about the extent to which this new technology would fundamentally transform many areas of society (Burnett and Marshall 2003). More sophisticated theories have subsequently emerged from the wreckage, arguing strongly against the tendency to technological determinism. In the academic realm, Kling and McKim (1999, 2000) argued that the impact of technology was highly specific to the discipline or even field concerned and would continue to be influenced by factors such as their individual needs for information sharing. Although Garrett, Lundgren and Nantz (2000) see the primary determining factor for Web use being individual faculty computer skills rather than discipline, the two are related and the disciplinary difference thesis is supported by empirical studies, one of general Internet use (Lazinger, Bar-Ilan & Peritz 1999), one of perceptions of academics (Herring 2001), and two centred on the phenomenon of Web linking (Tang and Thelwall 2003; Thelwall, Vaughan, Cothey, Li and Smith 2003). Hyperlinks between Web sites are an attractive data source because they indicate that information at one location is at least known about at another, showing a personal, cognitive or other connection between the authors or their work (Björneborn 2001).

The first Web link study to focus upon disciplinary differences compared the extent to which US departments of chemistry, psychology and history used the Web (Tang and Thelwall 2003). A very large disparity was found in both Web site sizes and interlinking, with chemistry making most use of the Web and history the least, with the difference being several orders of magnitude. A second study classified Web sites in Australia and Taiwan with high inlink counts, finding computing and some other subjects to be heavily represented but that the social sciences were much less prominent and philosophy and ethics appeared to be completely invisible (Thelwall et al. 2003). This exercise experienced considerable difficulty with inter-classifier consistency. It also ignored the large numbers of links to general Web pages, such as university home pages, which are very common (Thelwall 2002b) and may be created for subject-specific reasons (Thelwall 2003b). Another limitation was its restriction just to the highest linked Web sites, which would ignore subjects with a large number of low impact sites.

So far only one study has taken a large-scale statistical approach to comparing disciplines with online visibility, as measured by links to a site (inlinks). This research compared counts of links to Taiwanese universities with their science and social science research, as measured by citation index data. It was found that science research correlated more significantly with inlinks than social science (Thelwall and Tang 2003). Given the previous research showing a high degree of correlation between inlinks and research productivity in a wide variety of contexts (Thelwall 2001a; Smith and Thelwall 2002; Thelwall 2002a; Thelwall and Harries 2003, 2004; Thelwall and Wilkinson 2003) this supports the greater use of the Web in science than social science.

Link analysis

Link analysis has become a widespread tool for analysing the Web, partly because there is no other widely available source of data from which user activities can be inferred. The logical first choice for many purposes is actually something else, Web server logs. These allow researchers to track the activities of individual users with some degree of certainty, and are capable of giving much more information about Web users than

hyperlinks. They have been used, for example, to track digital library use (Marek and Valauskas 2002). Nevertheless, the problem is that these are typically hidden or protected on a site and so are not available for public scrutiny. A solution to this problem has been proposed (Evans and Furnell 2003) but its implementation does not seem likely in the near future because of a lack of immediate payoff for server administrators.

Hyperlinks from the raw data have been used for many different purposes including Web information retrieval algorithms (Brin and Page 1998; Kleinberg 1999; Henzinger 2001), sociological studies of Web communities (Park, Barnett and Nam 2002; Garrido and Halavais 2003), and Web structure mining (Henzinger 2001; Kosala and Blockheel 2001). These approaches all seem to stem from bibliometrics through analogies with citations (Rousseau 1997), with the explicit or implicit assumption that if one page links to another then either there is some relationship between the pages' contents or authors, or that the target page is more likely to be useful by dint of being targeted by a link. One strand of information science research, operating under the umbrella term Webometrics, has typically taken the latter perspective, seeing counts of links to a page, Web site or other space as being a potential indicator of its online impact (Almind and Ingwersen 1997; Ingwersen 1998; Björneborn and Ingwersen 2001). Subsequent research attempted to validate this assumption for sets of university departments (Thomas and Willett 2000; Chu et al. 2002; Li, Thelwall, Musgrove & Wilkinson 2003; Tang & Thelwall 2003), sets of universities within a country (Thelwall 2001a; Smith and Thelwall 2002; Thelwall and Tang 2003; Thelwall 2002a) and for academic journal Web sites (Smith 1999; Harter and Ford 2000; Vaughan and Thelwall 2003), eventually finding significant evidence in all cases. These studies all sought evidence of impact through correlations with existing measures of research quality or impact, yet relatively few links originating in a university site actually target formal scholarly publications such as online journal articles, although over 90% do target scholarly material of some kind (Wilkinson et al. 2003). As a result, inter-university links measure an amalgamation of a wide range of types of informal scholarly communication. In other words, a high impact university site is one that tends to provide useful scholarly information of a wide variety of types and/or whose scholars collaborate widely with other institutions.

There is a range of generic factors that affect the creation of links other than target page information value. Statistical physics models of Web growth support the notion that pages attract links simply because they are already highly linked to (Barabási and Albert 1999), although this tendency varies by type of page (Pennock et al. 2003). A common sense explanation for this would be that highly linked to pages are more 'visible' to Web users, either by following the links or by using link-influenced search engines (Arasu et al. 2001). Links can be seen as advertising in this sense (Thelwall 1999). From the same Web growth model it could be expected that older Web pages tend to attract more links, and this is borne out by evidence (Vaughan and Thelwall 2003). Links can also be created for offline reasons: because of relationships between page owners, or previous affiliations of the link page creator, for instance to link to the home page of the university where they obtained a Ph.D. (Thelwall 2003b). Another real world impingement is geography: pages in nearby universities are more likely to interlink than those in remote ones, at least in the UK (Thelwall 2002c). Finally, links can be created and replicated automatically, for instance automatically inserted 'created by' links in older Web authoring programs. All of these examples show the need to exercise caution when interpreting link counts. They should not be used as a primary information source in an evaluative role, but can be used for relational analyses and for supporting information (Thelwall 2002e).

Visibility, impact and influence

The terms 'visibility' (Vreeland 2000; Chu, He and Thelwall 2002), and 'impact' (Ingwersen 1998) have both been used as general descriptors for site inlink counts. The term 'impact' is commonly used to refer to that which is measured by citation counts, and so it is logical to use this also for inlink counts. The logic for citations is that they often represent use of a particular work by others (Merton 1973; Cronin 1984; Borgman and Furner 2002), and so imply a degree of impact within the scholarly community. This analogy only partially transfers to the Web because target page contents

rarely influence the broad contents of the source page (Wilkinson et al. 2003). Nevertheless, links do represent knowledge of at least the location (URL) of the target page, and 'impact' does seem a reasonable term for this, without getting into a semiotic discussion. Of course in both cases (articles and Web pages) impact is possible without links, for example simply by being read. Similarly, visibility is also an acceptable description since inlinks both often indicate that a page has been found, and make it more likely to be indexed by search engines (Vaughan and Thelwall 2004) and ranked higher in them (Brin and Page 1998). Again, there are other ways of finding pages than by following links or using search engines so visibility is also an imperfect term.

It will be necessary to distinguish between the evident impact of a subject and its impact influence, a distinction made here for the first time. *Evident impact* in the Web comes from links targeted at a page with evident subject content or ownership. This would include personal pages of faculty as well as online journal articles and teaching pages. *Impact influence* comes from links targeted at a page either without a subject affiliation, such as university home pages or recreational society pages, or at a page that has had its creation influenced or aided by subject specialists in any other way that is not manifest from its contents. This is an important distinction to make because a university Web site may be highly visible because of the activities of computer scientists, say, but not in a way that is apparent from studying page contents.

Research design

In order to address the research question of whether all areas of scholarship have a significant Web presence the concepts 'areas of scholarship' and 'Web presence' need to be made more concrete. The research design is opportunistic in the sense of using naturally occurring data for both of these but is nonetheless coherent. The U.K. is chosen for this study as the country with the most detailed published research information, the Research Assessment Exercise (RAE) (http://www.rae.ac.uk/). This information is reported in 68 subject categories from 'clinical laboratory sciences' to 'sports-related subjects' but, as will be explained later, this is too many to investigate for Web presence and so the five broad categories used by the RAE umbrella panels will be used instead.

I Medical and Biological Sciences
II Physical Sciences and Engineering
III Social Sciences
IV Area Studies and Languages
V Humanities and Arts

It is difficult to quantify Web presence. Ideally, the extent to which each area produces Web pages that are used by other universities would constitute its Web presence. The emphasis here is on users from outside of the particular university because Web publishing for internal institutional use is of a different character to that which outsiders judge valuable. Inter-university links will be used because links within a single university site are typically used for navigation purposes and are therefore not indicators of impact (Ingwersen 1998; Thelwall 2001a). As discussed above, links typically represent a range of types of informal scholarly communication and can be used to assess the online impact of research in a wide sense. As a practical step, only links between universities within the UK will be considered. There are thousands of universities in the world, so this is a practical necessity. There does not seem to be a significant difference between international and national link based impact (Thelwall 2002d) and so this is not a serious problem.

The next part of the research design is the method for determining a relationship between national Web presences, as measured by inter-university link counts, and the five areas of scholarship. A natural approach would be to classify a random sample of pages targeted by inter-site links. The difficulty of this task (Thelwall et al. 2003) leads to an alternative choice, the use of multivariate statistical techniques to seek a relationship between the variables.

Statistical techniques are not necessarily reliable on the data set because of potential relationships between the independent variables. In other words there may be *patterns* in subject specialties across universities: universities that specialise in one type of research may tend to specialise in another, completely unrelated type. For example, the oldest universities may tend to be strong in

both classics and medicine. Some type of triangulation is therefore necessary to support any findings. This will be based upon classifying a random sample of link target pages, using the same five categories above. If the two approaches give similar results then this would strengthen the findings. Note that the page categorization will take into account evident impact alone, whereas the modelling approach is also capable of incorporating impact influence.

Methods

Data sources

Data concerning the interlinking of academic institutions in the UK was obtained from a publicly available database created in June–July 2002 (http://cybermetrics.wlv.ac.uk). The database was compiled using a specialist information science Web crawler, designed for accurate coverage (Thelwall 2001b). The 110 institutions are those in the *Times Higher Education Supplement* list for students (Mayfield University Consultants 2000) but with the addition of the UK's largest distance learning institution, the Open University. The list includes the colleges of the federal universities of London and Wales as separate institutions, although legally they are not. This is consistent with the original list, public perceptions and Web site naming conventions.

The research data comes from the UK 2001 Research Assessment Exercise Web site (http://www.rae.ac.uk/). For each of the 110 institutions, this site gives a breakdown of the number of active researchers and the average quality of their research in the 68 different subject categories. The quality judgement was made by a panel of subject experts based upon the best four publications of each submitted researcher. On an international scale this is probably the most detailed and reliable source of information about academic research (McNay 2003; Adams 2002). The 68 subjects, called Units of Assessment (UoAs), are numbered 1 to 69 with 12 missing, and are grouped together into the five meta categories listed above to allow boards to compare the results of similar subjects. Note that this data source removes the need for any kind of classification exercise since researchers effectively self-classify their work by choosing the UoA to submit to. The UoAs themselves are defined and described by subject experts with the objective of grouping together research that could be assessed by a single team. Although the scheme does not conform to international codes (e.g. UNESCO 1997), the organic clustering approach gives a modern set of subjects with cognitive similarity and so it is ideal for our purposes.

Data analysis

Multiple linear regression will be used to build a model to 'explain' university inlinks in terms of the extent of research conducted in each of the five broad categories (subject research quantities are the dependent variables). It would be possible to assign a variable to each of the 68 subjects, but this would not allow effective tests for the goodness of fit because there are too few universities to give any power to the procedure. As a result we use only one variable for each of the five meta-categories instead. For each of the 68 categories in each university we estimate the total research conducted by multiplying the number of researchers by their subject score on a scale of 1 to 7. This conflates research quality with quantity but is a standard approach for RAE data (Mayfield University Consultants 2000; Education Guardian 2002; Thelwall 2001a). For each of the five areas the research productivity is the sum of all these values across all subjects in the area. The end result of this is a set of five numbers for each institution that estimate its research productivity in each category.

Before using the data we transform all variables (including link counts) by dividing them by the faculty numbers in the institution. This is necessary to factor out university size as a variable, because larger universities tend to have different research profiles than the smaller ones (Thelwall 2003a), which would otherwise skew the results.

The link data itself includes links between different universities only, as discussed above. Also, we do not count links between pages, only between domains. The purpose of this is to greatly reduce the impact of anomalous sources of multiple links. This approach is known as the *domain* Alternative Document Model (ADM) and is now a standard webometric technique (Thelwall 2002a; Thelwall and Harries 2003; Thelwall and Wilkinson 2003; Tang and Thelwall 2003; Thelwall and Tang 2003).

In order to apply multiple linear regression to a data set the variables involved should be normally distributed, but the variables obtained for research and link counts were not, being skewed. This was a major problem for the data analysis because the techniques for fitting regression equations require normality both to fit effectively and report the significance of the results. The data can be transformed to normal by taking square roots but fitting an equation to the transformed data would not be intuitively meaningful: the effect of the different types of research should be cumulative without any transformation. In order to resolve this difficulty with the data for the correct model being incompatible with the existing regression fitting techniques, we will fit both the transformed and untransformed data, a method triangulation (Tashakkori and Teddlie 1998). If the transformed model gives a similar importance to the variables as the untransformed model then we will use this as partial evidence to support the correctness of the first. The Enter method for linear regression was used because all disciplines create Web pages to some extent and so all should be represented in an accurate model.

Preliminary analysis revealed that two universities were persistent outliers, having more links than expected. These were dropped from the analysis because the counting methodology is not immune from variations due to administrative issues for Web hosting, particularly strategies for the issuing of domain names. The two in questions were Heriot-Watt and the University of Exeter. These two were ignored from all the statistical analyses but left in all of the graphs. In fact removing these two had little effect on the numbers involved and no effect on the conclusions of this paper.

Page classification

A random sample of 400 links between pages in separate universities was extracted using a specially written program operating on the same UK 2002 database. The selection was a genuinely random collection from the whole set, in contrast to previous similar exercises that sampled the same number from each university (Wilkinson et al. 2003) or the highest targeted pages (Thelwall et al. 2003). Standard link counting between pages was used (i.e. Page ADM) instead of the domain ADM to give an additional degree of triangulation: document model variation.

Each page was classified by the first author into one of the five general subject categories using the descriptions of the individual 68 categories. In most cases this was a straightforward choice with problems of subject content identification mostly falling within a category, and it was not thought necessary to crosscheck these with a second classifier. Note that subject-based educational initiatives for higher education teaching were classified by subject rather than as education. These are typically run by subject specialists rather than education researchers.

Table 1. Statistics for the untransformed variables

Group	Group name	Kolmogorov-Smirnov p	Spearman Correlation	Pearson Correlation
-All-	Overall	0.000	0.813	0.844
I	Medical and Biological Sciences	0.002	0.702	0.606
II	Physical Sciences and Engineering	0.001	0.780	0.775
III	Social Sciences	0.075	0.666	0.501
IV	Area Studies and Languages	0.007	0.602	0.514
V	Humanities and Arts	0.001	0.374	0.318

Results

Model 1: Untransformed variables

Table 1 gives the results of the Kolmogorov-Smirnov normality tests and reports Spearman and Pearson correlations between inlinks to whole universities and each of the five broad subject areas. Spearman values are more reliable since the data is significantly non-normal. The Kolmogorov-Smirnov p-value for university inlinks is 0.026. Research areas correlate highly significantly with inlink counts, but to greatly different amounts.

Table 2 reports the regression line fitted. The R-value is 0.874, only a slight improvement on the 'Overall' Pearson value of 8.44, and R-square is 0.764. The equation is not counter-intuitive: all coefficients are positive and a constant term is plausible since a university may attract links simply because it exists and not because of its actions (Pennock et al. 2002). Note that the significance values are only estimates since the data is not normal. The most remarkable part of the results

Table 2. Linear regression equation for the untransformed variables

	Coefficients	Significance
Const	0.237	0.000
I	0.219	0.000
II	0.565	0.000
III	0.172	0.002
IV	0.022	0.905
V	0.429	0.000

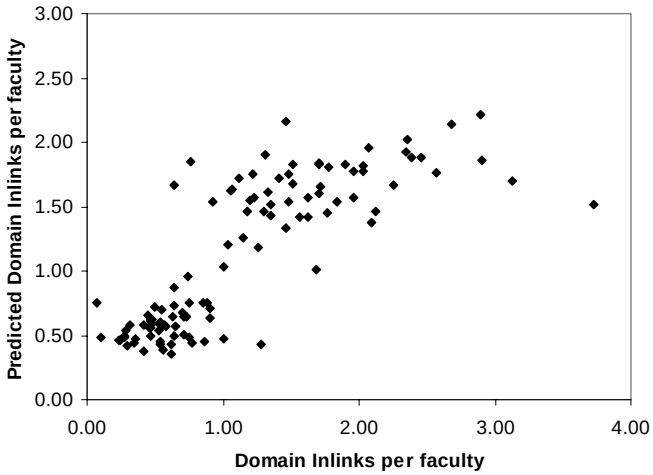

Fig 1. Linear regression model results (including outliers)

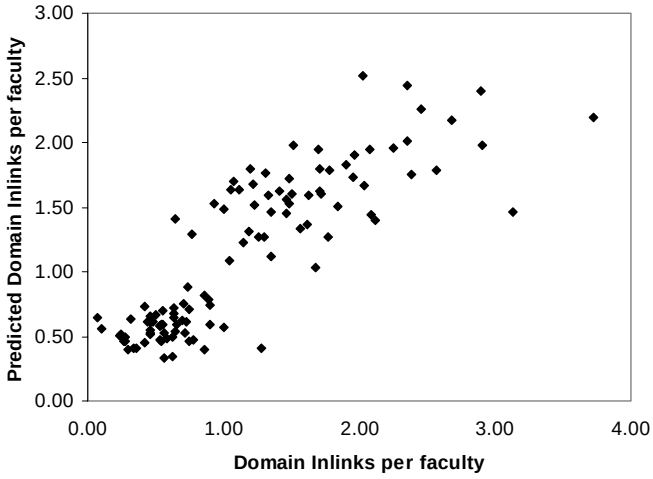

Fig 2. Multiple linear regression model results (including outliers)

is that Area Studies and Languages appears to have little impact on the Web, whereas Humanities and Arts has more despite its lower correlation with inlink counts.

Figures 1 and 2 show a standard linear regression based on overall research productivity (top) and the new equation. Clearly the more sophisticated multiple linear regression model makes only a very small improvement.

Table 3. Statistics for the transformed variables

Group	Group name	Kolmogorov-Smirnov p	Spearman Correlation	Pearson Correlation
-All-	Overall	0.002	0.813	0.841
I	Medical and Biological Sciences	0.075	0.702	0.679
II	Physical Sciences and Engineering	0.325	0.780	0.787
III	Social Sciences	0.878	0.666	0.596
IV	Area Studies and Languages	0.154	0.602	0.548
V	Humanities and Arts	0.174	0.374	0.278

Table 4. Linear regression equation for the transformed variables

	Coefficients	Significance
Const	0.329	0.000
I	0.169	0.003
II	0.414	0.000
III	0.114	0.077
IV	-0.003	0.973
V	0.286	0.000

Model 2: Transformed variables

The same calculations were performed on the square roots of all the data used. As can be seen from Table 3, the resulting variables are normally distributed. Interestingly, the total research productivity variable, which is not used in the model, is not normal – in fact it has a very bipolar distribution. The Kolmogorov-Smirnov value for the transformed university inlinks is 0.246.

As can be seen from Table 4, the results are broadly consistent with the first-reported model, with Area IV making an insignificant contribution to Web linking. The R-value is 0.861, only a slight improvement on the 'Overall' value of 0.841, and R-square is 0.741.

We tested for significant collinearity amongst the variables, but the results fell well short of the diagnostic figures suggested by Belsely, Kuh & Welch (1980) and so although all the variables correlate with each other, this does not occur to the extent that it threatens the results.

Page classification

In order to assess the relative sizes of each research area, Figure 3 shows the total quantity of

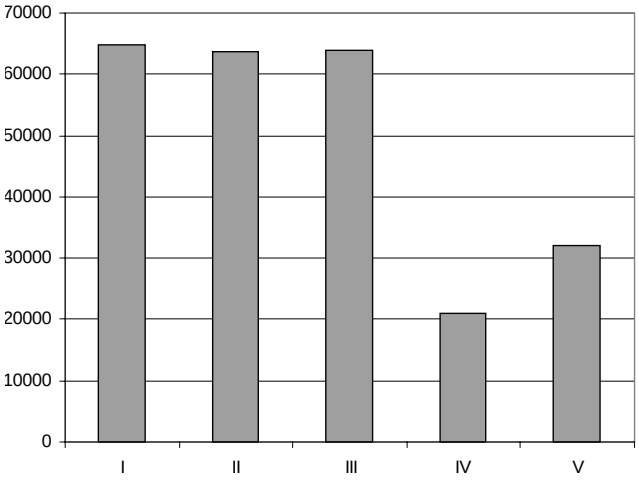

Fig 3. Relative size of research in each category.

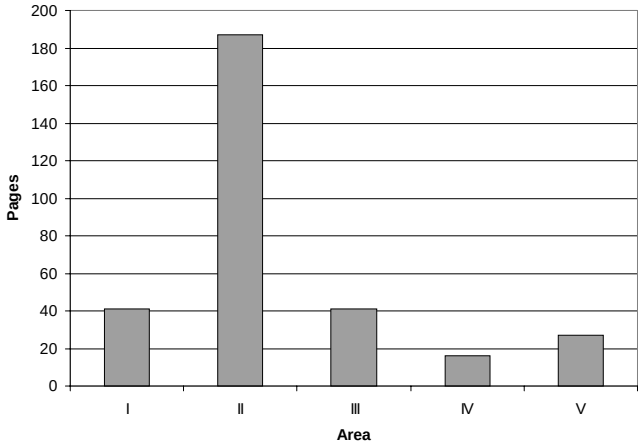

Fig 4. Subject area classification of 312 random link targets

research in each of the five categories. The vertical axis scale is nominal, calculated as described in the methods section. It is the sum of all RAE scores associated with all researchers in the area (rather than simply the total number of researchers). Figure 4 gives the results of the page classification exercise. This should be compared with Figure 3 to take into account the relative sizes of the subjects themselves. There were 83 further links that were classified as general, and 5 pages could not be found to be classified and were also missing from the Internet Archive. These are a possible source of bias in the results since Archive coverage is biased by link counts (Thelwall and Vaughan 2004), but since the numbers are small, this should not be a real problem: Figure 4 is unlikely to be significantly affected.

Discussion

Modelling

The results suggest that all areas of research make some contribution to the Web impact of the host university except for Area Studies and Languages. Physical Sciences and Engineering is the main contributor, as would be expected from previous research (Tang and Thelwall 2003; Thelwall et al. 2003) and the fact that it includes Computer Science. It is not dominant to the extent that it completely eclipses all other areas, as can be seen by the lower correlation coefficient between this type of research and inlink counts compared to the correlation between all research and inlink counts, although the difference is not large. The conclusion is also supported by the multiple linear regression results, which do not give Physical Sciences and Engineering a dominant role.

The results should be treated with a degree of caution because universities are not randomly generated and so there will be patterns in their choice of research specialties, due to tradition and perhaps such factors or influences as top universities dominating medical research. Nevertheless, UK higher education has self-consciously promoted diversity, even to the extent that many of the institutions did not carry out any research in one or more of the five areas. It could be this diversity that has allowed the multiple regression technique to single out Area IV as not contributing whilst Area V does, despite its low overall correlation with inlinks.

Collinearity is present in some degree between the variables, and in the nature of multiple regression it is difficult to separate out the effects of the different variables. It may be the case that the lack of institutions specialising in area IV has lead it to be difficult to differentiate from the other areas. Nevertheless, statistical collinearity tests were negative, indicating that the variables did make significantly different contributions and so this should not have been a problem.

Triangulation: Page classification

In the page classification exercise there were many pages not given a subject classification. These were mainly university home pages, but also included library and museum pages as well as local

information, recreational activities and student support services. The exact extent to which each of these was influenced by subject activities would be difficult to tell in many cases. For example, a link to a university home page could arise as a result of a collaborative research project (Thelwall 2003b), and a canoe club page could be created by a computer scientist with the necessary expertise; impact influence that the modelling technique could incorporate but not the page classification. Consequently, the results are not conclusive proof of the totality of subject impact on the Web.

Comparing figures 3 and 4 it is clear that area II (Physical Sciences and Engineering) has a disproportionate impact on the Web, confirming the modelling results, but hardly a surprising finding. The impacts of I (Medical and Biological Sciences) and III (Social Sciences) are approximately the same, as would be expected from their similar sizes (Fig. 3). Compared to I, III and IV (Area Studies and Languages), however, V (Humanities and Arts) has a high impact relative to its size. This is surprising given a previous finding that History in the US was almost invisible on the Web (Tang and Thelwall 2003). An examination of the classification for V revealed many pages with a focus on electronic librarianship and digital archiving. This could explain its relatively good showing. Nevertheless, a wide range of pages without a natural online project focus was present, including "Department of Music" and "Medieval drama links".

Another surprise is the comparison between I (Medical and Biological Sciences) and III (Social Sciences). Based upon Chemistry having many more links than Psychology, a previous study speculated that this would be true for all sciences compared to all social sciences (Tang and Thelwall 2003). But the gap between I and III is small or non-existent according to these results. Perhaps Medical and Biological Sciences use the Web less than other sciences in general.

There is a significant discrepancy between the results of the two approaches, principally with Area Studies and Languages being the target of more links in the classification exercise than would be expected from the statistical approach. This cannot be explained by the impact influence/evident impact dichotomy, because this could only be used to explain lower than expected page classification totals. The links are palpably there and so the most likely conclusion is that the regression results are not reliable. This may be due to too little variation institutional accommodation of area IV, although this did not show up as significant collinearity in the tests. Note that Area Studies and Languages may have relatively higher impact on an international scale, from the nature of their subjects, but this should affect both techniques equally.

A limitation of the study is that it covers the UK alone. Research cultures and traditions of Web use differ significantly on an international scale and so whilst it seems likely that the results would broadly apply internationally, this is not a necessary conclusion and individual national research councils would need to conduct their own research to verify these conclusions. In addition, the numbers for the page classification are too small to reach firm conclusions about the relative size of areas I, III, IV and V. Finally, the broad categorisation approach does not rule out the possibility that individual subjects in any of the five areas have very little Web presence.

Conclusion

The results show that there are clear differences in Web use by subject area, although not entirely of the kind previously suggested. Physical Sciences and Engineering is indisputably dominant but Humanities and Arts makes a surprisingly strong showing for its size. This is due in part to electronic librarianship and online archiving initiatives. Another unexpected result was the similarity in impact between the scientific Medical and Biological Sciences and the non-science Social Sciences. It was not clear whether Area Studies and Languages made a significant contribution to university Web presences, but the page classification exercise suggested that it did.

There are two possible explanations for the subject differences found. First, the Kling & McKim (2000) hypothesis that the extent and type of electronic communication needs varies between and within disciplines could be used to argue that the results are to be expected and are not a cause for concern. Second, the high showing of II and V in particular can be taken as evidence of the promise of the Web, a potential that the other areas are not reaching. It seems likely that the truth lies in a combination of the two, but the

answer cannot be obtained by the quantitative approaches used here. A constructivist approach is now needed (Tashakkori and Teddlie 1998) to find out how successful Web users are using it, why they are successful and if the lessons can be transferred to other disciplines. Given the centrality of the Internet for scholarly communication of many kinds, this is now an imperative task. In such an exercise, particular attention should be given to II and V in the search for exemplars. In terms of practical implications, research-funding bodies should consider whether action needs to be taken to ensure that opportunities are not being missed in the lower Web impact areas.

Acknowledgement

This research supported by a grant from the Common Basis for Science, Technology and Innovation Indicators part of the Improving Human Research Potential specific programme of the Fifth Framework for Research and Technological Development of the European Commission. It is part of the WISER project (Web indicators for scientific, technological and innovation research) (Contract HPV2-CT-2002-00015).

References

Adams, J. 2002. Research assessment in the UK. *Science* 296: 805.

Almind, T. C. and P. Ingwersen. 1997. Informetric analyses on the World Wide Web: Methodological approaches to 'Webometrics'. *Journal of Documentation* 53(4): 404–26.

Arasu, A., J. Cho, H. Garcia-Molina, A. Paepcke and Raghavan, S. 2001. *Searching the Web, ACM Transactions on Internet Technology* 1(1): 2–43.

Barabási, A. L. and R. Albert. 1999. The emergence of scaling in random networks. *Science* 286: 509–12.

Belsely, D. A., E. Kuh and R. E. Welch. 1980. *Regression Diagnostics: Identifying Influential Data and sources of collinearity*. New York: John Wiley and Sons.

Björneborn, L. and P. Ingwersen. 2001. Perspectives of Webometrics. *Scientometrics* 50(1): 65–82.

Björneborn, L. 2001. Small-world linkage and co-linkage. In: *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia* (pp. 133–134). New York: ACM Press.

Borgman, C. and J. Furner. 2002. Scholarly communication and bibliometrics. In: Cronin, B. (ed.): *Annual Review of Information Science and Technology 36*, Medford, NJ: Information Today Inc.: 3–72.

Brin, S. and L. Page. 1998. The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1–7): 107–17.

Brody, T., L. Carr and S. Harnad. 2002. Evidence of hypertext in the scholarly archive. In *Proceedings of ACM Hypertext 2002*: 74–75.

Burnett, R. and P. D. Marshall. 2003. *Web theory*. London: Routledge.

Chu, H., S. He, and M. Thelwall. 2002. Library and information science schools in Canada and USA: A Webometric perspective. *Journal of Education for Library and Information Science* 43(2): 110–25.

Cronin, B. 1984. *The citation process*. London: Taylor Graham.

Education Guardian. 2001. About the tables, http://education.guardian.co.uk [viewed 14 November 2003]

Evans, M.P., and S. M. Furnell. 2003. A model for monitoring and migrating Web resources. *Campus-Wide Information Systems* 20(1): 67–74.

Ferdig, R. E., and C. R. Hartshorne. 2002. Web and database network environments for educational supply and demand. *Campus-Wide Information Systems* 19(3): 92–98

Garrett, N.A., T. D. Lundgren, and K.S. Nantz. 2000. Faculty course use of the Internet. *Journal of Computer Information Systems* 41(1): 79–83.

Garrido, M. and A. Halavais. 2003. Mapping networks of support for the Zapatista movement: Applying social network analysis to study contemporary social movements. In: M. McCaughey and M. Ayers (Eds). *Cyberactivism: online activism in theory and practice*. New York: Routledge (pp. 165–184).

Harter, S. and C. Ford. 2000. Web-based analysis of e-journal impact: Approaches, problems, and issues. *Journal of the American Society for Information Science* 51(13): 1159–76.

Henzinger, M. 2001. Hyperlink analysis for the Web. *IEEE Internet Computing* 5(1): 45–50.

Herring, S.D. 2001. Using the World Wide Web for research: Are faculty satisfied? *Journal of Academic Librarianship* 27(3): 213–219.

Hyland, K. 2003. Self-citation and self-reference: credibility and promotion in academic publication. *Journal of the American Society for Information Science and Technology* 54(3): 251–59.

Ingwersen, P. 1998. The calculation of Web Impact Factors. *Journal of Documentation* 54(2): 236–43.

Kebede, G. 2002. The changing information needs of users in electronic information environments. *The Electronic Library* 20(1): 14–21.

Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5): 604–32.

Kling, R. and G. McKim. 1999. Scholarly communication and the continuum of electronic publishing. *Journal of the American Society for Information Science* 50(10): 890–906.

Kling, R. and G. McKim. 2000. Not Just a Matter of Time: Field Differences in the Shaping of Electronic

Media in Supporting Scientific Communication. *Journal of the American Society for Information Science* 51(14): 1306–20.

Kling, R. and E. Callahan. 2004, to appear. Electronic journals, the Internet, and scholarly communication. In Cronin, B. (Ed.) *Annual Review of Information Science and Technology* 37.

Kosala, R., and H. Blockeel. 2000. Web mining research: a survey. *ACM SIGKDD Explorations* 2(1): 1–15.

Lazinger, S. S., J. Bar-Ilan, and B. Peritz. 1999. Internet use by faculty members in various disciplines: a comparative case study. *Journal of the American Society for Information Science* 48(6): 508–18.

Li, X., M. Thelwall, P. Musgrove, and D. Wilkinson. 2003. The relationship between the links/Web Impact Factors of computer science departments in UK and their RAE ratings or research productivities in 2001. *Scientometrics* 57(2): 239–55.

Marek, K., and E. J. Valauskas. 2002. Web logs as indices of electronic journal use: Tools for identifying a 'classic' article. *Libri* 52(4): 220–30.

Mayfield University Consultants. 2000. League Tables 2000. *The Times Higher Education Supplement* April 14, II–III.

McAvinia, C. and M. Oliver M. 2002. "But my subject's different": a Web-based approach to supporting disciplinary lifelong learning skills. *Computers & Education* 38(1–3): 209–20.

McNay, I. 2003. Assessing the assessment: an analysis of the UK Research Assessment Exercise, 2001, and its outcomes, with special reference to research in education. *Science and Public Policy* 30(1): 47–54.

Merton, R. K. 1973. *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago Press.

Middleton, I., M. McConnell, and G. Davidson. 1999. Presenting a model for the structure and content of a university World Wide Web site. *Journal of Information Science* 25(3): 219–27.

Park, H. W., G. A. Barnett, and I. Nam. 2002. Hyperlink-affiliation network structure of top Web sites: Examining affiliates with hyperlink in Korea. *Journal of the American Society for Information Science* 53(7): 592–601.

Pennock, D. M., G. W. Flake, S. Lawrence, E. Glover, & C. L. Giles. 2002. Winners don't take all: Characterizing the competition for links on the Web. *Proceedings of the National Academy of Science (PNAS)* 99(8): 5207–5211.

Rees, M. 2002. Not worth the paper. *New Scientist* 2370: 27.

Rousseau, R. 1997. Sitations: an exploratory study. Cybermetrics, 1(1). URL: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html [viewed 14 November 2003]

Smith, A. G. and M. Thelwall. 2002. Web Impact Factors for Australasian universities. *Scientometrics* 54(3): 363–80.

Smith, A. G. 1999. A tale of two Web spaces: comparing sites using Web Impact Factors. *Journal of Documentation* 55(5): 577–92.

Tashakkori, A. and C. Teddlie. 1998. *Mixed methodology*. London: Sage.

Tang, R. and M. Thelwall. 2003. Disciplinary differences in US academic departmental Web site interlinking, *Library & Information Science Research* 25(4): 437-58.

Thelwall, M. 2001a. Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology* 52 (13): 1157–68.

Thelwall, M. 2001b. A Web crawler design for data mining. *Journal of Information Science* 27(5) 319–25.

Thelwall, M. 2002a. Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university Web sites. *Journal of the American Society for Information Science and Technology* 53(12): 995–1005.

Thelwall, M. 2002b. The top 100 linked pages on UK university Web sites: high inlink counts are not usually directly associated with quality scholarly content. *Journal of Information Science* 28(6): 485–93.

Thelwall, M. 2002c. Evidence for the existence of geographic trends in university Web site interlinking, *Journal of Documentation* 58(5): 563–74.

Thelwall, M. 2002d. A comparison of sources of links for academic Web Impact Factor calculations. *Journal of Documentation* 58(1): 60–72.

Thelwall, M. 2002e. Research dissemination and invocation on the Web. *Online Information Review* 26(6): 413–20.

Thelwall, M. 2003a. *Which way forward for UK corporate universities? An analysis of subject specialisms*. University of Wolverhampton.

Thelwall, M. 2003b. What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation, *Information research* 8(3): 151. URL: http://informationr.net/ir/8-3/paper151.html [Viewed 14 November 2003].

Thelwall, M. and G. Harries. 2004, to appear. Do better scholars' Web publications have significantly higher online impact? *Journal of the American Society for Information Science and Technology*.

Thelwall, M. and G. Harries. 2003. The connection between the research of a university and counts of links to its Web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. *Journal of the American Society for Information Science and Technology* 54(7): 594–602.

Thelwall, M. and R. Tang. 2003. Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan. *Scientometrics* 58(1): 153–79.

Thelwall, M., L. Vaughan, V. Cothey, X. Li, and A. Smith. 2003, to appear. Which academic subjects have most online impact? A pilot study and a new classification process. *Online Information Review* 27(5).

Thelwall, M. and L. Vaughan. 2004, to appear. A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research*.

Thelwall, M. and D. Wilkinson. 2003. Three target document range metrics for university Web sites. *Journal of the American Society for Information Science and Technology* 54(6): 489–96.

Thomas, O. and P. Willett. 2000. Webometric analysis of departments of librarianship and information science. *Journal of Information Science* 26(6): 421–28.

UNESCO. 1997. International Standard Classification of Education (ISCED 1997). URL: http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm [Viewed 14 November 2003].

Vaughan, L. and M. Thelwall. 2003. 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(1): 29–38.

Vaughan, L. and M. Thelwall. 2004, to appear. Search engine coverage bias: evidence and possible causes, *Information Processing & Management*.

Vreeland, R. C. 2000. Law libraries in hyperspace: A citation analysis of World Wide Web sites. *Law Library Journal* 92(1): 9–25.

Wang, P., M.W. Berry, and Y. Yang. 2003. Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology* 54(8): 743–58.

Wilkinson, D., G. Harries, M. Thelwall, and E. Price. 2003. Motivations for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication. *Journal of Information Science* 29(1): 59–66.

Editorial history:
paper received 7 July 2003;
accepted 30 September 2003.