ESRC End of Award Report

The World Wide Web of Science: Emerging Global Sources of Expertise

RES-160-25-0031

Ralph Schroeder, Alexandre Caldas, Shefali Virkar,
and William H. Dutton
Oxford Internet Institute, University of Oxford

Background

The World Wide Web of Science (WWWoS) project explored the degree to which the Internet is supporting a 'winner-take-all' pattern of access to scientific expertise. Rather than enabling expertise to be accessed from a more diversified array of sources, a winner-take-all, or power law, hypothesis posits that online access will reinforce central sources of science expertise, leading access to expertise to become more concentrated.

This role of the Internet in 'reconfiguring access' to information (Dutton 2004), including scientific expertise, is becoming increasingly important as people shift more time and attention to the use of online resources (mainly the Web) to access information and expertise in an expanding array of information sectors. While there has been little research in this area, the topic of search engines, online resources, and access to expertise has become a prominent research issue during the course of the present project.

Objectives

The aim of this research project was to assess whether and to what extent the Internet and the Web could be transforming access to sources of scientific expertise. There are competing theoretical perspectives, either for a concentration effect for the most heavily linked and accessed sites (driven by 'power laws' or a winner-takes-all effect) or for a greater diversification or 'democratization' of online resources.[1] This led to a number of related questions, such as whether the Internet is associated with the use of more global versus local sources of expertise? Are there differences across disciplines? How central a role is the Internet playing in shaping access to information and expertise among scientists?

In addition to these substantive questions, the project also sought to contribute to the development and application of Webmetric techniques, and their combination with in-depth interviews and focus groups. The project team found that this triangulation yielded greater insight into the role of the Internet in accessing different sources of expertise.

[1] This literature is reviewed in the project proposal, but the classic work remains Merton (1988). An important popularization of the 'winner-take-all' thesis is provided by Frank and Cook (1995), and moves into the context of the Internet with Barabási (2003).

Methods

This was a one-year exploratory research project. It employed multiple traits and methods to empirically examine patterns of access to scientific expertise. We approached this by describing networks of scientific communication. Since critical collaboration tends to be focused on particular areas of science, we studied communication patterns surrounding specific issues. This was accomplished by sampling key global issues that reflected a range of challenges of world-wide importance. Topics were chosen also to avoid favouring a winner-take-all hypothesis by selecting issues that were not inherently more concentrated, such as in a few centres for big science.

Given the exploratory nature of the proposed research, the original proposal identified six topics from a potentially wide range of issues. These were:

1. Climate Change,
2. Internet and Society,
3. Poverty,
4. Trade Reform,
5. Terrorism and
6. AIDS/HIV

The project was anchored in Webmetric analyses of the structure of networks of relationships on the Web within each of these six topic areas. This involved crawling the World Wide Web to identify and collect links which could then be analyzed to determine the structure of online networks in each issue area.

The Webmetric results were validated and extended through two complimentary methods. Validation of a sub-sample of issues provided added confidence in the findings across all six issue areas:

- A set of in-depth interviews were conducted with a sample of known experts -- scientists working within four of the six issue areas. Two were anchored primarily in the social sciences (Internet and society, terrorism) and two in the natural sciences (climate change and HIV/AIDS). Four topics from the six were chosen to allow for greater in-depth analysis of the Web representations and more interviews with researchers within each topical area. The project team interviewed 20 researchers, five from within each of the four topics. Given limited funding for fieldwork, researchers were chosen from universities in Oxford and Greater London, and do not represent a random sample from a well defined population of experts. Primarily face-to-face, but also some telephone interviews, were semi-structured and asked a series of questions to contextualize and identify how these researchers use the Web and other offline and online resources. All the interviews were transcribed and coded using qualitative analysis software (e.g., NviVO) to extract relevant information from the text of interviews.

- The results of our Webmetric analyses were validated also by two expert panels, similar to focus groups. One looked at ‘climate change’ and another at

'Internet and society'. Three and five participants respectively were chosen based on accessibity to the research centre, with each focus group was conducted in Oxford, at the OII. These enabled us to compare lists of the top sites with the respondents' lists of top sites, providing a means for judging the face validity of the Webmetrics.

Changes in the Research Design

In-depth interviews replaced the originally proposed Web-based survey. The team's concern over problems with response rates and limited number of questions that are practical for a Web-based survey led to abandoning this method. The interviews provided richer information than was originally envisaged, supporting a number of publications.[2]

One further change from the original proposal was the decision not to undertake a study of Usenet groups as an additional source of evidence on academic networks. Usenet is one of the early systems, created in the 1970s, that supported distributed discussions over the Internet which has continued to enable users to post and read articles within a wide range of newsgroups. The analysis of Usenet groups across these areas was pursued at the first WWWoS workshop, and the OII was able to obtain permission and a license to analyse Microsoft's archive holdings. However, it was not taken beyond this point for two major reasons. First, a paper had just become available (Matzat 2004) which covered this topic well. Secondly, it became increasingly clear that the use of Usenet had become a niche area that was increasingly unrepresentative of online access by scientists generally. Its use was also being marginalized by the role of search engines and other online resources, such as digital libraries and online datasets. It was therefore decided to focus project resources on in-depth interviews with scientists to validate, interpret and extend the Webmetric results.

Results

A number of patterns and themes have emerged from this exploratory research, which are substantively and methodologically useful in shaping follow-on research. The details are developed further by the two nominated outputs, but are summarized in this section.

First, the central finding of our study is that the winner-take-all hypothesis fails to reflect the more complex structures of scientific networks of expertise. It is true that a small proportion of sites capture a disproportionate share of links, but the Webmetric analysis and interviews suggest that:

- The structure of networks was more fractal in structure than would be expected by a simple winner-take-all hypothesis;
- Numerous clusters of institutions and websites are more prominent than others, but there are many winners sharing the attention of more specialized networks of researchers and other users;

[2] Key early findings are provided by Fry et. al. (forthcoming), but additional publications are in progress.

- The type of search and the topic make a difference in the overall structure of expertise, i.e., whether it is a directed search on a topic where there are established sources, or if the topic requires a more exploratory approach aimed at tapping a range of diverse sources;
- Search engines, and Google in particular, play an increasingly important gate-keeping role – shaping winners and losers, though this function varies between the four topics investigated and the type of search engine, as different search engines yield different sources of expertise. Indeed, the study provided a robust validation of the heterogeneity of search engines, the qualitative differences among them in terms of which functionalities they provide, and the apparently different content and Web spaces they provide in indexing and search services.
- Researchers display significant differences in how they access online resources (for example, if they go from search to publications, or vice versa; go from online to offline resources, or vice versa; look for people or institutions, etc.), which mitigates potentially more systematic impacts of the Internet on who goes to what sources of expertise;
- There was a US bias in the search engine results, when compared to the local UK sources that our UK experts consult, though this was more the case where the research topic prioritised national online resources (HIV/AIDS public health sites) than for the other three topics.

Secondly, there is an embedded social structure inherent in the distribution and sharing of resources on the Web. Similar patterns of the centrality of key resources, their connectivity and the way they are clustered, emerge on the Web spaces of all six topics. These social networks can be seen as electronic social networks, and identifying such structures will be of value in designing the search engine technology of the future.

Thirdly, the qualitative findings of this exploratory research suggests the potential value of embedding in future search engine technology more ‘user oriented’ and topic specific functionalities, such as peculiar characteristics of ‘geography and locality’ relevant to the topic. Policy considerations influencing the distribution of resources as well as the more academic nature of certain kinds of information on the Web might explain significant differences in the social networks embedded within these electronic networks.

Methodologically, the study supported the value of using Webmetric analyses and interviews in a complementary way. For example, we were able to successfully compare the most heavily linked and highly rated sites for each topic with those that researchers regularly accessed and used. The following sections provide a more detailed discussion of findings within and across the two methodological approaches.

I. Results of the Webmetric Analysis

The core empirical data set was created from Webmetric research tools, which enabled us to crawl the Web for links to and from sites to analyse patterns of access on the World Wide Web to scientific and technical knowledge in the six broad global issue areas of scientific research: ‘Climate change’, ‘Poverty’,

'AIDS/HIV', 'Terrorism', 'Internet and society' and 'Trade reform'. We employed two complementary strategies of using Web search engines for accessing information and expertise within these six topical areas.

The first strategy was based upon a comparison of searches using 6 different search-engines (Google, Yahoo, MSNSearch, AskJeeves, Gigablast and Google Scholar). Using each search engine, we conducted a simple Web search for a set of 3 keywords for each topic.

The second and complementary strategy involved an extended search, using a 'meta-search-engine' which combined 31 search engines[3] and also used an extended set of keywords in each topic. Following from this meta-search strategy, a 'structurally embedded' analysis was conducted based upon the Webmetric information resulting from network analysis of Web linkages on the Web in the six global topics.

As noted above, in-depth interviews and expert focus groups were used to assess not only the Webmetric results within two of our six research topics but also to evaluate the two different Web search strategies. This expert assessment, combined with lists of URLs provided by researchers, gave us a good indication of the validity of the results of searches. Both reinforced our findings on the fractal nature of networks of scientific expertise, which made the value of the structurally embedded search strategy more evident.[4]

The Heterogeneity of Search-Engine Technology

Results for our six topics obtained from a simple strategy demonstrated the heterogeneity in *recall* capacity of search engines, as well as the heterogeneity of results across search engines. Different search engines yielded different results. However, similar patterns and outcomes were consistent across the range of topics (see Table below).

[3] The 31 search engines used for the meta-search strategy were: http://search.about.com/, http://www.alexa.com/, http://altavista.com/, http://search.aol.com/, http://askjeeves.com/, http://search.dmoz.org/, http://search.dogpile.com/, http://www.euroseek.com/, http://msxml.excite.com/, http://www.alltheWeb.com/, http://findwhat.com/, http://www.google.com/, http://hotbot.lycos.com/, http://search.jayde.com/, http://www.looksmart.com/, http://search.lycos.com/, http://www.mamma.com/, http://www.moonmist.info/, http://search.msn.com/, http://search.netscape.com/, http://srch.overture.com/, http://www.rolist.com/, http://www.scrubtheWeb.com/, http://www.search.com/, http://www.searchgate.co.uk/, http://www.searchhippo.com/, http://s.teoma.com/, http://search.thunderstone.com/, http://dpxml.Webcrawler.com/, http://www.wisenut.com/, http://search.yahoo.com/.

[4] An overview of the 'extended search' model methodology is available in Caldas et al (forthcoming).

Table – Heterogeneity of Search Engines regarding Recall Capacity

Search-Engines Results for Six Global Topics

	Google	Yahoo	MSNSearch	AskJeeves	Gigablast	Scholar.Google
Terrorism	51,000,000	103,000,000	11,030,218	7,185,000	73,936	86
HIV/AIDS	26,200,000	44,600,000	6,754,236	7,099,000	196,608	156,000
Climate Change	34,500,000	37,400,000	6,566,133	6,363,000	72,178	227,000
Internet and Society	1,230,000	4,430,000	763,500	313,100	57,757	8,710
Trade Reform	1,040,000	2,220,000	496,267	183,700	56,850	20,700
Poverty	142,000	348,000	80,337	35,600	15,742	113
Self-Reported Index Size	**8,168,684,336**	*	**5,000,000,000**	*	**2,024,193,536**	*

* Data not available

Search engines were queried in the period 8 - 10 August 2005, using the following Keywords for each of the six global topics

URLs: www.google.com; www.yahoo.com; search.msn.com; www.askjeeves.com; www.gigablast.com; scholar.google.com

Keywords:

Internet and Society: "Internet and society" OR "Internet research" OR "Internet studies"

Climate Change: "Climate change" OR "Global Warming" or "Ozone Depletion"

Terrorism: "Terrorism" OR "Terrorist organisation" OR "Terrorist network"

Trade Reform: "Trade reform" OR "Trade liberalisation" OR "Trade and development"

Poverty: "Poverty research" OR "Poverty statistics" OR "Poverty and globalisation"

HIV/AIDS: "HIV/AIDS" OR "HIV infection" OR "HIV prevention"

First, the ‘recall capacity’ of search engines varies significantly, i.e., the potential total number of results on any query coming from each search-engine is significantly different across Google or AskJeeves or MSNSearch. This suggests that search engines have different ‘sizes’. They ‘scale’ differently.

Secondly, there are differences in the quality and ‘genre’ of results across the search engines. In this regard, Google, MSNSearch and AskJeeves form a distinct cluster. They are specialised on ‘general search’ without much focus but with very-large scale capabilities. A second distinct group is represented by Yahoo, which employs a different kind of search mechanism. It is characterised by being a ‘directory’ organised by themes and topics, and providing directory-oriented search engine functionalities. A third group is formed by Gigablast, a more specialised search engine, more focused on scientific and technical information, and providing wider search functionalities, which is focused on ‘document-based’ search and connectivity among documents on the Web. Finally, Google scholar constitutes a completely distinct search engine, only focused on indexing scientific and technical documents. In some respects, Google Scholar is the realisation of the full potential of Gigablast, although Gigablast is restricted to document search.

Thirdly, we can see that there are important differences across the search engines in terms of the ‘content’ indexed in the various crawlers and indexing technologies. This is evident from a detailed analysis of the ‘outcomes’ of the 30 first results for each topic across the search engines, showing the ‘diversity’ in resulting content.

The Structure of Knowledge on the Web

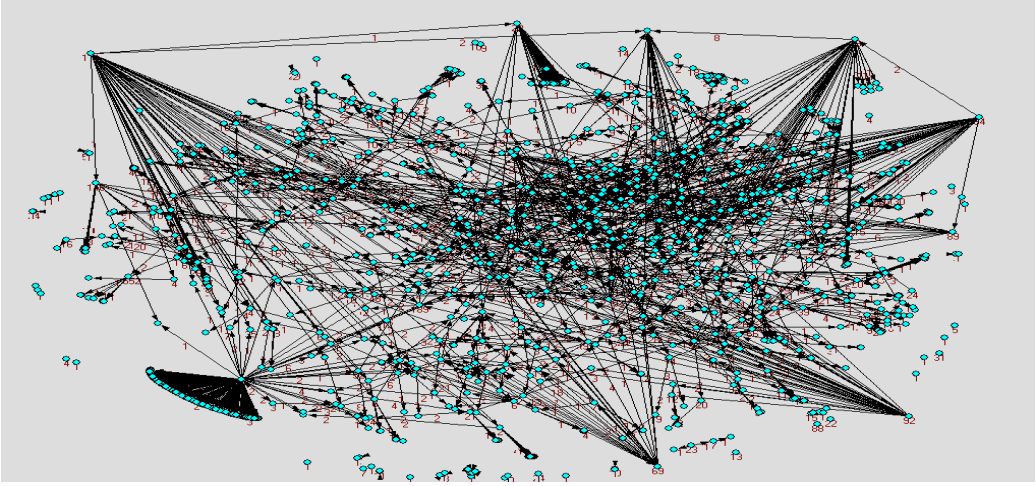
Against this background of the overall differences between search engine results, we turned to whether the above results were linked to an underlying

structure of links on the Web. This can be done by examining the results from the 'meta-search', combined with Webmetric link analysis for our six topics.

The results confirm regularities across the whole set of topic areas, with regard to centrality, connectivity and the subgroup structure of these 'Web networks'. Thus they provide support for the hypothesis that 'social networks' are embedded in the distribution of Web resources, with important implications for search engine functionality and access to knowledge on the Web.

A 'Web space graph' for 150 selected URLs in 'Climate change' was calculated by examining the *inlinks* and *outlinks* on the Web. A total of 3,489 distinct nodes was defined through this process. A graphical representation of the structure of this Web community is provided in the following Figure for 'Climate Change'.

Figure – Web space graph of 150 URLs in Climate Change



Maps for the other five topics ('Internet and society', 'poverty', 'HIV/AIDS', 'terrorism' and 'trade reform') are being made available on our Web site[5], revealing similar patterns of centrality and connectivity.

These patterns point to a 'fractal structure', where nodes in the network form hubs and (several) spokes connecting to these 'hub' nodes. This basic structural pattern is replicated across each of the six areas. The overall network then appears to form a 'fractal' where each 'subgroup' is interlinked by a reduced number of connections to other more distant groups. Even if there are differences across the six topics, the same basic underlying 'abstract' structure of a fractal can be consistently identified across these Web spaces.

The detailed analysis of these Web inter-linkages provides additional support to the hypothesis that there are some similar patterns within these electronic networks, in terms of centrality of resources, connectivity and subgroup structure.

[5] See: http://www.oii.ox.ac.uk/research/project.cfm?id=22

For example, the indicators for average distance among reachable pairs of nodes are relatively small (2.2 and 2.3 respectively), which suggest that despite the significant size of these Web networks (3,489 nodes and 20,839 arcs in 'Climate change', and 3,815 nodes and 31,736 arcs in 'Internet and Society'), there is nevertheless a short distance interconnecting any two different nodes on the Web.

Despite the size and evolution of these systems, there is a significant process of 'clustering' within these electronic networks. For instance, there are sub-groups of tightly knitted and highly clustered nodes within these larger networks. Factors explaining the relatively high 'clustering' coefficients cannot be easily explained by any technical bias, and require further exploration of offline characteristics to understand the dynamics of each cluster.

Finally, it is noteworthy that these networks exhibit low density and a 'sparse' nature. In any of the six Web networks there are a large number of nodes that are loosely connected or even isolated, and a much smaller number of nodes in a centrally connected and more highly interconnected 'core'. Overall, links on these six topics form a connected network of very low density on the Web.
In short, the 'fractal' nature of these Web networks could point to 'democratization' (low density) and 'reinforcement' (clique) effects rather than to a 'power law/winner-take-all' effect.

The On-line and Off-Line World

These Web networks are assumed to be reflections of underlying 'social networks' among scientists. Traditional offline 'social networks' are key to understanding and predicting the online inter-linkages on the Web.
In fact, the combination of the extensive data derived from the Webmetric analyses with the qualitative validation piloted by the two expert panels on 'Internet and Society' and 'Climate Change', tends to validate the assumption that 'offline social structures' are key drivers for the distribution of online resources and Web access to information. For instance, several of the researchers in the expert group on 'Climate change' referenced the importance of geography and locality in the patterns of access and distribution of resources in their field (particularly because climate change varies significantly according to geography and peculiarities of locality in different places of the world). According to their own reflections, this has direct implications for their reliance on alternative Web resources.

Climate change policy was also referenced as a key factor explaining the distribution of Web resources. The comparison between Europe and the US in climate change policy are a case in point. Significant differences in priorities in climate change policy understandably explain differences in the overall emphasis put on the delivery of information, services and other resources online.

The 'topic specificity' of resources on the Web particularly the distinction between 'academic networks' as opposed to a more 'societal view' on access to Web resources, was an interesting point brought out by several of the experts.

There was some agreement that significant differences might accrue from more academic or scientifically-oriented Web spaces, as opposed to wider and more broadly defined spaces of Web information and resources. This was several times referenced in the expert group of 'Internet and Society', but also in the community of climate change researchers.

II. Interviews: A Validation of Webmetric Analyses

In order to help interpret and validate the results of our Webmetric analyses on patterns of access to information, we conducted a series of semi-structured interviews in a sub-sample of the original six global domains. These were Terrorism, HIV/AIDS, Climate Change, and Internet and Society. In total twenty UK-based active researchers were interviewed; five from each of the four domains.

Interviewees were asked about their research background, key institutions, groups and people in their research networks, and the variety of online resources they used. Questions also focused on their online search strategies, such as the tools they used for finding information, the keywords they used and what kind of entities they tended to search for e.g. people, groups or institutions. The interviews also covered direct validation of specific aspects of the Webmetric data. For this we asked respondents to comment on the Google representation of key institutions, organizations and people in their domain and the extent to which it overlapped with their own mental model of the domain e.g. their individual perception of what constitutes the core set of resources and sources (it is important to note that this is different to a particular mental model they may have at any one time in relation to a situational information need). The Google representation was derived during the webmetric data gathering and analysis by retrieving the top-thirty URLs in each domain from Google.com based on the following keywords:

Climate Change	'Climate change', 'Global warming', 'Ozone depletion'
Internet and Society	'Internet and society', 'Internet research', 'Internet Studies'
HIV/AIDS	'HIV/AIDS', 'HIV Infection', 'HIV prevention'
Terrorism	'Terrorism', 'Terrorist organisation', 'Terrorist network'

Any overlaps or inconsistencies between the Google representation of each domain and the participants' own mental model was further validated by coding the websites, institutions, organizations, people and other resources they reported using throughout the interview transcripts and then comparing this list with their responses to the Google representation. The interviews were recorded, transcribed in full and analysed using the Nvivo software for qualitative data analysis.

Geographic Orientation of Domains

The interviews revealed that there was only a limited overlap between the Google representation and researchers' own mental models of key networks, structures and organizations. Researchers reported that many of the key online

resources in their domain were missing from the Google representation. The extent of the overlap appears to be domain dependent, with those researchers working within a more nationally orientated information environment reporting less of an overlap. For example, the HIV/AIDS researchers reported using national journals, national charity organizations, national statistics, and national public sector organizations, but none of these appear in the top thirty Google results for generic domain keywords (even when the search was repeated using Google.co.uk). Climate Change researchers, on the other hand, for whom the geographical boundaries of research were far more 'international', were able to recognise many more Google results on the Climate Change validation sheet.

Networks of Excellence

In addition to the gaps that participants identified in response to the Google representation, a number of institutions, people and journals which researchers named as being leading academic institutions, people and other key organisations/resources in their field, did not appear in the top thirty Google results. This was particularly true for academic institutions and leading researchers named by the interviewees. In the cases where participants recognised some of the top sites from the list, or named key institutions, groups or people that did appear in the top thirty results, those identified were unlikely to appear in the top 10 results.

Web-based Search Strategies

Respondents recognized and noted that search engines, such as Google, were blunt instruments with respect to their ability to pin point information being sought. This was apparent from searches within fields of their own expertise. Also, the UK-based researchers perceived a persistent US bias in the results of their searches. Nevertheless, search engines, and Google in particular, became and remained the main tool for finding sources and resources on the web.

Though there were similarities in web-search strategies across each of the four domains, there were also important differences. For example, while respondents reported using Google almost to the exclusion of all other generalist Internet search engines, the role that it played in their wider information environments varied considerably. In the HIV/AIDS and Internet and Society domains, for instance, Google was mainly used as what Beauvisage (see van Couvering 2006) calls an 'aide memoir', a tool for locating known sources. In contrast, for researchers of Terrorism, Google played a relatively more central role in exploring the object of research and identifying relevant sources. This may be due to the amorphous, shadowy nature of the subject matter itself – websites of terrorist groups and the message-boards, chatrooms and blogs associated with them are constantly being shut down by national intelligence agencies, only to resurface with new web-addresses, and the only way to locate these and other sources like them is for researchers to 'excavate' resources across a range of domain boundaries.

One possible explanation for differential domain patterns in the role of Google and other Internet search engines as information seeking tools could be the

extent to which important documents are scattered across domain boundaries. The consequence of this for web searching is that in low scatter fields, resources and sources can be found using a clearly circumscribed set of keywords and are likely to be produced by a limited number of dominant gatekeepers. Of the case studies, HIV/AIDS was the domain with the least scatter and this could explain why Google was used more as an 'aide memoir' than as an exploratory tool. Terrorism and Climate Change researchers on the other hand described their domains as scattered in terms of resources and respondents reported using Google for finding diverse sources more than in the other two domains.

The Role of Gatekeepers

The characteristics and role of the predominant gatekeepers varied across each of the four domains. The interview responses indicated a differentiated shift towards the decentralization of gatekeepers on the Internet. For example, in climate change, 'hybrid research centres' produce and disseminate important sources; and policy or academic research centres are key producers of information sources in Internet and society research. In HIV/AIDS research, in contrast, although non-profit organizations were key producers and disseminators of information and played an important gatekeeping role, traditional gatekeepers such as publishers still maintain a central position in the information environment because of the continued importance of peer-reviewed articles disseminated through discipline-centric aggregated databases such as *PubMed*.

The information environment of the Terrorism researchers was similar to that of the HIV/AIDS researchers in that, while non-governmental and not-for-profit organizations play a central role in disseminating primary information resources, publishers still had an enduring role as gatekeepers to academic research. In terrorism, dissemination of research via books plays a major role in the scholarly communication system and still remains closely interrelated to the recognition and reward system. Research in terrorism is of a sensitive nature, which may account to some extent for the sustained importance of the traditional gatekeepers such as publishers. In contrast, the gatekeepers in the information environments of the climate change and the internet and society researchers were more decentralised. This meant that - rather than access to information being coordinated by a predominant gatekeeper - there were multiple gatekeepers providing specific resources in niche areas.

In summary, as with quantitative webmetric results, our qualitative interview findings show that there is no uniform 'winner-takes-all' effect in the use of online resources. Instead, there are different kinds of gatekeepers for the four topics we examined and for the types of information that are sought. It is therefore important not just to identify a concentration or democratization effect, but rather to refine under what circumstances the search for expertise will be dominated by certain results and exhibit biases, and when, instead, researchers will be led to the resources they seek and to a variety of results. Particular characteristics of a domain's information environment will determine whether

Google and other Internet search engines function as a *facilitator* or as an *influential gatekeeper*.

Research Activities

This section describes a range of activities undertaken during the course of this project, which complemented the ongoing research. While any list would be partial, the key activities included:

1. An early workshop with Dr Marc Smith of Microsoft Research in Redmond was held at the OII on May 19, 2005, specifically to identify issues related to online expertise and how to map this research.
2. A special lecture took place at the OII with Alexander Macgillivray, Senior Product and Intellectual Property Counsel, Google, on November 3, 2005, which was attended by a wide audience from outside the University.
3. A two day workshop was held from February 9-10 2006 with top scholars in the field, entitled, 'The World Wide Web and Access to Knowledge'. The workshop included several international specialists, including: Mike Thelwall, Wolverhampton, 'Analysing Public Debates of Science through Blogs and News Sources'; Elizabeth van Couvering, LSE, 'Strategies for Gatekeeping the Internet: the Power of Search'; Andrea Scharnhorst, Royal Netherlands Academy of Arts and Sciences, 'Indicators from the web - making the invisible visible?'; and Matthew Hindman, Arizona State University, 'From Production to Filtering: The World Wide Web and Shifting Patterns of Information Exclusivity'. There was also a presentation by the project itself, 'The World Wide Web of Science and Global Expertise: Democratizing Access to Knowledge?' by Ralph Schroeder, Jenny Fry and Shefali Virkar, OII. This workshop was useful in presenting a state-of-the-art overview and situating the project in ongoing research.
4. An additional workshop held June 26, 2006, called 'Researching the Web of Knowledge'. It featured three expert speakers from around the world: Franz Barkjak, University of Applied Sciences Northwestern Switzerland, 'Hyperlinks in Academia'; Rob Ackland, Australian National University, 'Virtual Observatory for the Study of Online Networks'; and Michael Macy, Cornell University, 'The Cornell Internet Database: 200 Terabytes for Social Research' (about a new web archiving project). All three talks provided evidence of the complementary work that is starting to be carried out around the world, the increasingly systematic approaches to how people search for information and access expertise, electronically, and the growing interest in methods to study search processes, a topic which is being extended in new directions.
5. Attendance and participation in a variety of seminars and conferences (including AoIR), as illustrated in the outputs.

Outputs of this Research Project

The project has yielded a number of working papers, conference presentations and publications, including:

Schroeder, R., Caldas, A., Mesch, G., and Dutton, W. (2005), 'The World Wide Web of Science: Reconfiguring Access to Information'. A paper and presentation delivered at the First International Conference on e-Social Science, Manchester, June 22-24 2005, The paper and slides are available online at : http://www.ncess.ac.uk/events/conference/2005/papers/

Caldas, A., Schroeder, R. Mesch, G. and Dutton, W. (forthcoming), 'Patterns of Information Search and Access on the World Wide Web', Special Issue on 'The Social, Political, Economic and Cultural Dimensions of Search Engines', *Journal of Computer-Mediated Communication*, edited by Eszter Hargittai (forthcoming, accepted for publication, in draft form).

Fry, J., Virkar, S. and Schroeder, R. (2006), 'Search Engines and Expertise about Global Issues: Well-defined Territory or Undomesticated Wilderness?', in Michael Zimmer and Amanda Spink (eds). Web Search: Interdisciplinary Perspectives (forthcoming, accepted for publication, draft form).

> This paper will be presented at the Association of Internet Researchers (AoIR), Sept. 27-30, 2006, in Brisbane, Australia.

The following two publications also benefited from work on this project:

Caldas, A. (forthcoming), 'The Winner Structures It All Hypothesis', *Research Policy.* Forthcoming. Caldas used the techniques of the project in relation to university rankings, extending the project to a domain which will be useful to researchers and university administrators.

Fry, J. (2006). 'Google's Privacy Responsibilities at Home and Abroad', *Journal of Librarianship and Information Science* (forthcoming, July issue). Fry analyzed how the controversy over the US government request for Google queries and the uses and limitations of Google in China highlight issues of access to expertise.

Copyright allowing, the outputs will be available and advertised on the project website: < http://www.oii.ox.ac.uk/research/project.cfm?id=22>. Future publications evolving out of this research will also be made available on this site.

* Given the exploratory nature of the Webmetric analyses, based on a limited sample of issues, and sites relevant to each, we do not anticipate a need to place the Webmetric data within the UK Data Archive. Anyone wishing to replicate our work could obtain the datasets from our OII Website, or more likely, generate new datasets with a different sample of issues and sites. This would provide a more powerful replication of our exploratory work. Moreover, we are able to provide the Data Archive with the Webmetric results, if requested.

Impacts of the Research

The significance of the topic addressed by this research project has literally 'taken off' since the project's launch, with growing media interest in the matter of how people search for information. For example, the *Economist* specifically covered the topic of the 'winner-take-all' effect on the Web, focusing on the ideas of one of the project workshop participants, Matthew Hindman[6]. This may have been one of the major implications of our research for public understanding.

In addition, the principal investigator, Ralph Schroeder, was interviewed by BBC Radio 4 (Chris Bowlby chris.bowlby@bbc.co.uk and Diane Coyle) for the 'Analysis' programme specifically about the outcomes of the project and in relation to Websearch and topics such as the influence of Google. The programme was aired on July 27 and 30, 2006.

Another consequence of research was a discussion with Stephen Schifferes and his colleagues at the BBC Online, an organization which is currently acutely concerned with the reach and scope of its online presence and visibility. Another group that was interested in the results were the Climate Change group at Oxford University, which has had considerable success in engaging users online in the concerns about global warming and found the results important in developing its online strategy. <http://www.eci.ox.ac.uk/climatesystems.html>

Across all these examples, ESRC research findings have been used to inform public debate and to highlight key questions of policy for industry and government. The question of access to and 'visibility' of certain sites are certain to be the subject of continuing debates with which members of the research team will engage.

Future Research Priorities

As mentioned earlier, interest in this field of research has expanded dramatically over the past year. It has become evident from this project, but also from a wide range of other studies, that online resources are becoming an increasingly central aspect of how people seek information and expertise. This is true for academic research, as indicated by the present study, but also for the general public.[7]

This exploratory research supports the value of extending work on this topic to develop larger samples from more researchers across more domains, and to follow trends over time. There remains only a limited amount of research about how academics and other users actually use the web. The use of Webmetrics within this project demonstrates the value of this new technique for expanding our understanding of this field, and has also led the OII to consider building

[6] *The Economist*, November 19, 2005, p. 97.
[7] See, for example, Dutton, W., di Gennaro, C., and Millwood Hargrave, A. (2005), *The Internet in Britain: The Oxford Internet Survey (OxIS*). Oxford: Oxford Internet Institute.

greater capacity for webmetric analyses within the OII and in other areas of research.

There is also scope for studies in specific areas. For example, the principal investigator of this project (Schroeder) together with OII colleagues is drafting a research proposal provisionally entitled 'The World Wide Web of News: The Global Scope and Reach of the BBC's Online Services'. There have been several discussions with BBC online staff about the project which would have major policy implications for the BBC (to what extent should it aim at a global audience, especially as the BBC is the most widely used source of news in the UK and perhaps globally, but it is also funded primarily by UK taxpayers). This project would extend the target population of the 'World Wide Web of Science' project to include a global public instead of academic researchers, and shift the focus from global research topics to political information and debate.

The 'World Wide Web of Science' project has crystallized some key parameters of this topic (use of Google, gate-keeping, shift to offline resources) for a small but important group. Future research should focus and expand on:

- search for information by wider publics (for example, journalists, groups interest in political information, school children) and with a wider range of topics
- the dominance and gate-keeping function of Google and other search tools
- refining methods (how to combine quantitative and qualitative methods is still a major issue, which is the subject of much current discussion)
- understanding the effect of a migration to online resources and away from offline resources, overall and in relation to specific areas
- the 'visibility' or 'web presence' of different topics. This visibility issue is not just an issue of access to expertise, but also a policy issue: is some content 'marginalized' by different search strategies?

Compared to the rapidly growing significance of these topics, this is still a vastly under-researched area. The question posed by this small project demands a much more systematic project in the future, and it can be hoped that this research project, in addition to having answered some questions, has also led to a refinement and useful extension of the research agenda in this rapidly changing field.

References

Barabási, A-L. (2003), *Linked*. Cambridge, Massachusetts: Perseus Publishing.

Caldas, A., Schroeder, R. Mesch, G. and Dutton, W. (forthcoming), 'Patterns of Information Search and Access on the World Wide Web', *Journal of Computer-Mediated Communication* (forthcoming, accepted for publication).

Dutton, W. H. (2005), 'The Internet and Social Transformation: Reconfiguring Access', pp. 375-397 in Dutton, W. H., Kahin, B., O'Callaghan, R., and Wyckoff, A. W. (2005), *Transforming Enterprise*. Cambridge: MIT Press.

Frank, R. H., Cook, P. J. (1995), *The Winner-Take-All Society: Why the Few at the Top Get So Much More Than the Rest of Us*. New York: The Free Press.

Fry, J., Virkar, S. and Schroeder, R. (2006), 'Search Engines and Expertise about Global Issues: Well-defined Territory or Undomesticated Wilderness?', pp. forthcoming in Zimmer, M. and Spink, A. (eds). Web Search (forthcoming).

Matzat, U. (2004), Academic Communication and Internet Discussion Groups: transfer of information or creation of social contacts?, *Social Networks,* 26: 221-55.

Merton, R. (1988), 'The Matthew Effect in Science', *Isis*, 79: 606-623.

van Couvering, E. (2006), Web Behaviour: Search Engines in Context. Available at http://personal.lse.ac.uk/VANCOUVE

(c. 5000 words excluding tables and supporting material)