Web Issue Analysis: An Integrated Water Resource Management Case Study

Mike Thelwall[1]
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478
Katie Vann: Royal Netherlands Academy of Arts and Science. POB 95110, 1090 HC Amsterdam. The Netherlands. Email: Katie.Vann@niwi.knaw.nl
Tel: +31 20 462 8681Fax: +31 20 665 8013
Ruth Fairclough
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: r.fairclough@wlv.ac.uk
Tel: +44 1902 321000 Fax: +44 1902 321478

In this paper web issue analysis is introduced as a new technique to investigate an issue as reflected on the web. The issue chosen, Integrated Water Resource Management (IWRM), is a United Nations-initiated paradigm for managing water resources in an international context, particularly in developing nations. In common with many international governmental initiatives, there is a considerable body of online information about it. 41,381 HTML pages and 28,735 PDF documents mentioning the issue were downloaded. A page URL and link analysis revealed the international and sectoral spread of IWRM. A noun and noun phrase occurrence analysis was used to reveal the issues most commonly discussed, revealing some unexpected topics such as "private sector" and "economic growth". Whilst the complexity of the methods required to produce meaningful statistics from the data are a disadvantage for its easy interpretation, it was still possible to produce data that could be subject to a reasonable intuitive interpretation. Hence web issue analysis is claimed to be a useful new technique for information science.

Introduction

Policy makers and researchers sometimes need to find out about the spread or diffusion of particular issues of concern. In response, some literature-based methods have been developed that allow issues to be investigated through the occurrences of certain words or phrases. The pioneering work of Lancaster and Lee (1985), for example, showed that it was possible to track the growth of an issue ("acid rain") in academic papers through the databases of the time. The inclusion of different types of database (e.g., academic and mass media) may give more general policy-relevant information concerning the shifting of issues between different spheres of interest. For instance, Wormell (2000) used multiple databases to track concepts related to the Danish Welfare State, with press coverage taken to represent public opinion. The web seems to provide an environment in which issues can be analysed across many different areas of influence, due to the wide variety of organisations and individuals that publish online. Two disadvantages of the web compared to many literature databases, however, are that it can be difficult to conduct analyses of events after they occur, since web pages naturally die or are replaced (Koehler, 2004), and also the web is not organised in a way that makes it easy to retrieve relevant high quality data. There is some interesting current web-based information-centred research that tracks topics on the web from the perspective of the effectiveness of the web as an information source, rather than to understand the underlying dynamics of the issue (Bar-Ilan, 2000; Bar-Ilan & Peritz, 1999, 2004). The methods used will not be suitable for retrospective issue analyses because they rely upon the collection of time series data from search engines for a pre-defined search phrase.

[1] *This is a preprint of an article to be published in the Journal of the American Society for Information Science and Technology © copyright 2004 John Wiley & Sons, Inc.* http://www.interscience.wiley.com/

Many researchers have developed new methods to analyse web-based phenomena. Site-based *link analysis* approaches often investigate collections of web sites through their interlinking (e.g., Bar-Ilan, 2004a; Björneborn, 2004; Garrido & Halavais, 2003; Heimeriks, Hörlesberger, & van den Besselaar, 2003; Ingwersen, 1998; Musgrove, Binns, Page-Kennedy, & Thelwall, 2003; Park, 2003; Polanco, Boudourides, Besagni, & Roche, 2001; Rogers, 2002; Thelwall, 2004a). These methods are most effective for collections of web sites that naturally interlink. The interlinking of individual pages has also been investigated to identify 'communities' of pages (Flake, Lawrence, Giles, & Coetzee, 2002; Thelwall, 2003) and (in conjunction with text analyses) collections of topic-relevant pages (Bharat & Henzinger, 1998; Chakrabarti, Joshi, Punera, & Pennock, 2002; Kleinberg, 1999), but neither of these tasks produce significant information about the context of web issues. Other site-based approaches analyse the set of links to or from a collection of web sites, rather than just the links between sites in the set (Chu, 2005; Foot, Schneider, Dougherty, Xenos, & Larsen, 2003; Thelwall, 2004b; Thelwall & Aguillo, 2003; Vaughan & Hysen, 2002; Vaughan & Wu, 2004). Link analysis methods can be used to generate information about relationships between sites or collections of pages on the web but a generic limitation of all link methods is that link creation is a voluntary activity and the absence of a link does not indicate the absence of a connection. Nevertheless, link analyses have been able to give useful exploratory results in many of the above examples. Note, however, that the site-based approach of previous large-scale link analyses is not suited to the investigation of online issues, particularly those attracting governmental interest. This is because government and NGO sites can be very large and can include multiple unrelated issues. For example two consecutive government pages may be legislative acts on unrelated concerns that were approved at the same time. Hence for an issue analysis, individual related pages need to be processed rather than whole web sites.

A few web *text analysis* studies have investigated the text of web sites from an issue perspective, either using content analysis (Weare & Lin, 2000) or metrics based upon word frequency counting (Price & Thelwall, 2005; Thelwall, 2004c, 2005). Whilst content analyses can give detailed and insightful information, they are labour-intensive and time-consuming. Word frequency analyses are less labour-intensive but suffer from polysemy: they join together words that have different meanings. They also suffer from partial information, for example the phrase "University of Wolverhampton" looses significant meaning when split into its individual words. Computational linguists have addressed similar problems through algorithms that can accurately identify the meanings of words in text and extract coherent text units, such as phrases (Mitkov, 2003). Such techniques have been applied to the web for linguistic purposes such as identifying patterns of language use (Mair, 2003) and automatic translation (Meyer, Grabowski, Han, Mantzouranis, & Moses, 2003), but not to analyse contemporary political issues on a large scale. Text-based methods have been used to map issues in large sets of documents, but not using natural language processing techniques (Leydesdorff, 1989; Leydesdorff & Hellsten, 2005, to appear).

Our new method, *web issue analysis*, uses both text and page-based link analysis, and incorporates some natural language processing algorithms from computational linguistics. Our intuition is that the text of a related web page may convey the essence of the page more fully than could be inferred from the page inlinks or outlinks. In particular, a large-scale text-based investigation into related web pages may reveal the aspects of the issue that are most commonly discussed, and which other topics are most closely related to it. Link analyses may give different types of findings, perhaps regarding the flow of information or the relationships between organisations.

In this paper, web issue analysis is assessed through a full-scale case study of a relevant issue: Integrated Water Resources Management (IWRM).

Integrated Water Resource Management

The Aral Sea has shrunk in size by over 50% since the 1950s as a result of farmers near its feeder rivers diverting water to irrigate their land (http://www.dfd.dlr.de/app/land/aralsee/back_info.html). The sea and surrounding land have now become highly saline and affected by significant pesticide residues. This change has been catastrophic for the livelihoods of some 100,000 people in Uzbekistan and Kazakstan, particularly local fishermen and farmers. Like many other environmental problems, this is an international issue because the farmers benefiting from the diverted river water are in different countries to the Aral Sea communities. Problems like this have lead the United Nations (UN) to develop Integrated Water Resource Management (IWRM), a paradigm for the management of water resources on an international scale (Newson, 1999). IWRM was instigated at the 1992 United Nations Conference on Environment and Development (UNCED) Dublin-Rio conference on water and became part of Agenda 21, a wide-ranging UN agreement for sustainable development (United_Nations, 1992). Now, over a decade later, it is important to assess how this issue is perceived on an international scale, and who is discussing it.

There is a considerable body of published work that assesses and describes IWRM. The UN and various individual nations have sponsored many non-governmental organisations (NGOs) that promote and assess various aspects of IWRM. These NGOs produce reports from their own perspective and often publish significant resources online (e.g., Cap-Net, http://cap-net.org). There are also academic publications describing IWRM from a critical-pedagogical perspective (e.g., Blatter & Ingram, 2001; Calder, 1999; Maganga, Kiwasila, Juma, & Butterworth, 2004; Newson, 1999; Thomas & Durham, 2003) and research into individual case studies from a management perspective (Alfarra, 2004; Levesque, 2001), some of which underpin masters level IWRM courses. In addition, there is some scientific research into solution methods for individual IWRM problems (Wright, 2003) and there are also several academic reports of global IWRM initiatives (Bonell, 2004). What is also needed, however, is an independent overview of the current status of IWRM, taking an international, holistic perspective. This can help stakeholders such as the UN to decide upon future policy directions, and can help NGOs involved in individual projects to set their work in a contemporary international context. One previous qualitative study has given a web-based IWRM perspective on a small-scale, with the main finding that traditionally disenfranchised groups, such as local communities in developing nations, appear to be absent from the web (Thelwall, Barlow, & Vann, 2005).

Method

The web issue analysis research design is to retrieve web *pages* related to IWRM and then to extract indicative information about IWRM through the text and links of these pages. The case study objective is to explore the key current issues and the international spread of IWRM through a large-scale analysis of web documents. Clearly this approach is self-limiting, able only to assess aspects of IWRM that are reflected online. As such, it is likely to be biased towards academic, official and developed nation sources (Introna & Nissenbaum, 2000; Vaughan & Thelwall, 2004) and away from certain key types of activities, such as communications to and from affected peoples in developing nations (Thelwall, Barlow et al., 2005). Nevertheless, an advantage is that widespread international coverage may be obtained without dependence upon the cooperation of stakeholders and interest groups. A technical advantage is that all web documents are in electronic form and hence easily available for an automated analysis. Web issue analysis is designed to give a descriptive analysis of issues on the web. Its findings can be taken (a) as a partial, exploratory study of all IWRM (both online and offline) in order to provide hypotheses for further study using non-web sources, or (b) at face value as descriptions of the "web issue". The former is the goal of this case study: the issue rather than the web itself is the object of interest. As discussed above, this is only partially possible because of biases inherent in web sources. In practice, the focus on the issue rather than the web manifestation of the issue takes the form of extensive *data cleansing* in order to remove from the data wherever possible all occurrences of web phenomena that serve to obscure IWRM. For example, in our analysis, replicated links between sites, such as on a site-wide navigation bar, are not allowed to dominate the link count data because these probably reflect a single decision to create a link, even though the links are replicated through the navigation bar because of a web design decision.

Data Collection

Figure 1 summarises the main stages of the data collection and processing, which was conducted in March 2005. After the first activity, computer programs conduct all subsequent stages separately. The tasks are described in detail below.

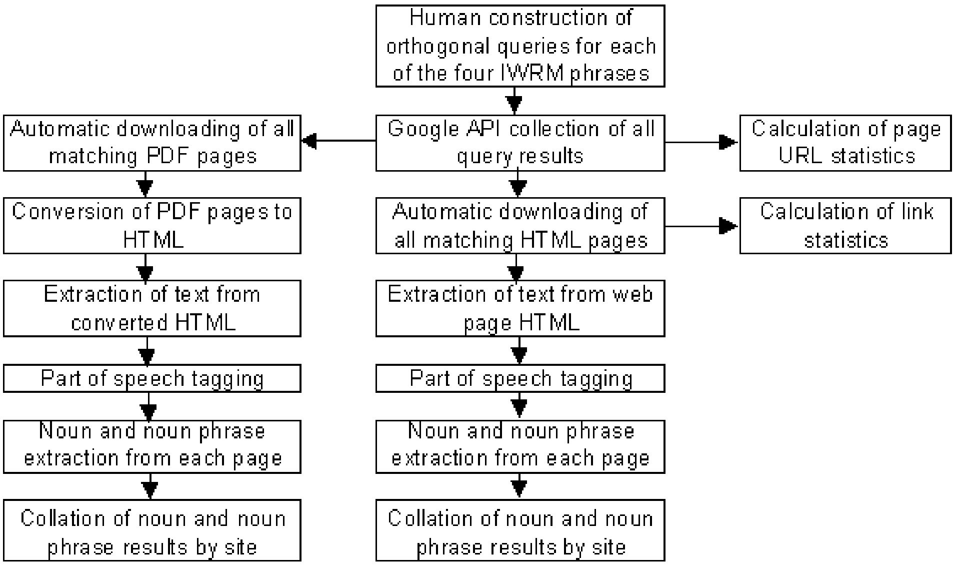
Human construction of orthogonal queries for each of the four IWRM phrases
Automatic downloading of all matching PDF pages
Google API collection of all query results
Calculation of page URL statistics
Conversion of PDF pages to HTML
Automatic downloading of all matching HTML pages
Calculation of link statistics
Extraction of text from converted HTML
Extraction of text from web page HTML
Part of speech tagging
Part of speech tagging
Noun and noun phrase extraction from each page
Noun and noun phrase extraction from each page
Collation of noun and noun phrase results by site
Collation of noun and noun phrase results by site

Fig. 1. Main data collection and processing stages.

We operationalise *IWRM-related web pages* as web pages that either mention IWRM or "Agenda 21" and water. The only practical method to identify a reasonably comprehensive collection of IWRM-related web pages is to use a commercial search engine, because the web is too big for coverage by a private crawler unless the study is limited to a specific set of sites (Thelwall, Vaughan, & Björneborn, 2005). Since there were too many pages to manually collect from a search engine, the choice was restricted to Google because it allows automated queries via the Google API (www.google.com/apis/), an interface that programmers can use to access Google's database. Others have previously used this tool for research purposes (Cimiano & Staab, 2004). Unfortunately, the Google API tends to return significantly less results than the standard Google interface, and so the API results represent a subset of Google, which itself only covers a part of the web (e.g., Lawrence & Giles, 1999), an unavoidable limitation. The queries below were formulated as likely to indicate IWRM-related web pages and unlikely to be present in irrelevant pages. Testing showed that the queries produced a vast majority of related pages, although they included some spurious matches (e.g., a few pages for the International Workshop on Resource Mobilisation, also known as IWRM). Note that an English bias is introduced by the word "water" in the second query.

- IWRM
- "Agenda 21" water
- "Integrated Water Resources Management"
- "Integrated Water Resource Management"

In principle, these queries could be submitted to Google to produce four lists of matching URLs and then the lists could be combined and duplicates eliminated. This was achieved after developing methods to overcome two obstacles.

First, the Google API permits only 1,000 queries per user per day, with 10 results per query. Hence the queries had to be run over several weeks in order to obtain a full set of matching URLs.

Second, the Google API returns a maximum of 1,000 URLs per search phrase (in 100 consecutive pages of 10 results). The search queries listed above returned many more than this. In order to retrieve a complete set of results, artificial non-overlapping searches were generated by adding or subtracting additional words to the original search phrases in order to get a large collection of more specific artificial queries that would individually return less than 1,000 matches, but would collectively match all of the URLs that the core phrases match. For example, the second query was translated into 1,024 (i.e. 2^{10}) sub-queries by adding or subtracting all 10 words from the list: “supply land climate sea environmental crisis toxic assess disease issue”, resulting in queries such as the following, where IWRM was eliminated from all queries because this word was searched for in the first set of queries instead (i.e. the IWRM queries).

- “Agenda 21” water supply land climate sea environmental crisis toxic assess disease issue -IWRM
- “Agenda 21” water supply environmental crisis disease -issue -assess -toxic -sea -climate -land –IWRM
- “Agenda 21” water -issue -disease -assess -toxic -crisis -environmental -sea -climate -land -supply -IWRM

Using all possible combinations of words and –words in the above example should give 1024 queries having individually less than 1,000 matches but which collectively combine to give results equivalent to the modified core query: "Agenda 21" water -IWRM. In practice, however, the Boolean logic of search engines is not perfect (Bar-Ilan, 2004b; Mettrop & Nieuwenhuysen, 2001; Smith, 1999) and so the above queries resulted in some overlaps and will probably also have missed some matching pages, but after eliminating duplicates the method was able to obtain a large number of matching URLs.

All URLs identified from the Google API were downloaded with the research crawler SocSciBot (HTML pages) (Thelwall, 2004a), or a purpose-built file downloader (PDF files). The small number of non-HTML, non-PDF files were ignored in the belief that they would be unlikely to contain information that was not in HTML or PDF form in other documents. A possible exception to this is PowerPoint presentations, but there were only a small number of these.

Note that no information is collected on links *to* IWRM-related web pages. Although Google could have been used for this, it would have required over 70,000 queries. Using the Google API’s standard rate of 1000 queries per day, this would have taken a minimum of 70 days and was judged impractical. Nevertheless, it is likely that the majority of links to IWRM-related pages are themselves in IWRM-related web pages and would hence be included in the sample.

Data Processing

Data cleansing and counting methods

Data cleansing is a key issue because without it the results would give little useful information about IWRM. This is due to a number of reasons that cause the raw data to be dominated by factors unrelated to IWRM, such as web site design choices. To illustrate this, in the raw data the 8th top linked page from the data set was http://www.visitblackpool.com because 422 pages on a Blackpool local council site linked to this tourist site on their standard links bar (which also mentioned Agenda 21 and water). After the data cleansing described in the paragraph below, these 422 links did not dominate the results because they counted as one “site link” as a result of originating from a single web site.

The choice of types of links to include is important. Web sites are often designed as coherent collections of pages to be easily navigated by users and hence the most common types of links within a site are links to other pages within the same site such as the home page. Web sites are also often highly repetitive in terms of both links and text. For example, some sites have the same navigation bar on all pages, perhaps including a few links to other sites, and some standard text, such as contact information, the organisation name or a disclaimer. Thus, to avoid the results of a large-scale analysis being dominated by large, repetitive sites, where possible counts are made on a ‘per site’ basis rather than per page. For example a link replicated over 1,000 pages of a site would count as one site link rather than 1,000 (page) links. Table 1 below summarises the counting method for each type of data reported.

During the analysis it was discovered that over 200 pages were in the results that were created from the open source directory dmoz.org's wastewater category list. Since replicated pages are undesirable for the analyses, they were filtered out by excluding all source pages with URLs containing the strings "water_resources/wastewater" or "water_resources%2fwastewater".

URL Analysis

The purpose of the URL analysis is to describe the overall distribution of the URLs of IWRM-related web pages. URL distributions can be summarised through a lexical analysis, counting the number of pages associated with each domain name and part of domain name. The following aggregation units are used.

- Domain name: Normally the section of an URL after the http:// and before the first subsequent slash (for example, www.wlv.ac.uk from http://www.wlv.ac.uk/index.html).
- Top-level domain (TLD): The section of the domain name following the final dot (uk in the above example).
- Second/Top-level domain (STLD): For domains using an official second-level domain name structure, the section of the domain name following the penultimate dot, otherwise the TLD (.ac.uk in the above example).
- Site: The section of the domain name including the STLD and one additional (dot-separated) preceding segment of the domain name (wlv.ac.uk in the above example).

Examples of the above definitions can be seen in the results tables below. The rationale for lexical URL analysis is that domain names are often reasonable indicators of the origins of pages. For example, pages with domain name ending in .ac.uk mostly originate in UK universities and those with domain name ending .wlv.ac.uk mostly originate in Wolverhampton University. Domain names are limited, however, because the widely used com domain is a poor indicator of content, and there are other similar discrepancies. Hence lexical URL statistics must be interpreted as indicative rather than definitive.

Link URL Analysis

Link URLs in IWRM-related HTML pages may give useful information about the most commonly linked-to web sites, organisations and countries. These links may reflect many causes, including the provision of relevant information interorganisational connections, or authority for a topic (Bar-Ilan, 2005). Link URLs are summarised with the same lexical method as page URLs above, and with the same limitations.

Two types of link URLs were excluded from the analysis. Site self-links are links between pages within a single web site. Web sites commonly carry a navigation bar on each page, resulting in extensive site self-linking. This type of link is a less valuable indicator of content relationships than inter-site links because the latter presumably represent a more considered choice. Hence all site self-links are excluded from the link analysis.

Some sites attract many links for very general reasons. For example there are large numbers of links to amazon.com referring to books and to adobe.com to download its document reader. Such links are peripheral to an issue analysis. The following list was therefore used to automatically exclude links with domain names containing the specified text: adobe.com; altavista; amazon; dmoz.org; doubleclick.net; google; Microsoft; w3.org; yahoo; netscape; statcounter.com; nedstatbasic.net; netscape. Note that these will match a range of sites. For instance, "google" matches all Google sites including www.google.com, directory.google.com, and www.google.co.uk.

Interlink Analysis

Relationships between organisations or their exchange of information may be reflected in some cases through links between their web sites, although recall that the absence of a link does not imply the absence of a connection. Inter-site links originating in IWRM-related pages many particularly reflect IWRM-related relationships. The purpose of the interlink analysis is to highlight such relationships. The same data filtering was used as for the link URL analysis, producing statistics about the interlinking of sites, based upon links found in the downloaded IWRM-related pages.

Noun analysis

The noun analysis is designed to discover the most common topics in IWRM-related web pages which was achieved in several stages. Each HTML web page in the data set was stripped of its HTML tags and reduced to its text using a simple filtering program. Each PDF document was converted into a HTML page using the program PDF Ripper and then converted to plain text using the same filtering program.

Some nouns in text can be directly identified through the use of a dictionary but many nouns are polysemous, such as 'egg', which could be a noun or a verb. Hence the task of automatically identifying nouns in text is highly non-trivial. It is part of the computational linguistics field of natural language processing and is typically tackled by using artificial intelligence techniques (e.g. Markov models) in conjunction with a dictionary. It can be best achieved through assigning 'parts of speech' to all of the words in a document (Brill, 1992), i.e. classifying words as nouns, verbs etc.

The part of speech tagger Lingua-EN-Tagger (http://search.cpan.org/dist/Lingua-EN-Tagger/) was used to tag predicted parts of speech for all the words in each text file and then to create a list of nouns for each file. These word lists were then merged for all of the pages with URLs sharing a common domain name. Finally, a combined noun list was created recording the number of domains in which each noun occurred in at least one page. The use of a domain frequency rather than a simpler page frequency measure is a new data cleansing technique based upon the alternative document model concept previously used in link analysis (Thelwall, 2002). It is designed to stop the frequencies of some nouns being artificially inflated by multiple occurrences of similar documents or parts of documents within a single domain. A high domain frequency count, then, is a strong indication of the importance of a noun for IWRM because many different web domains include it in one or more of their IWRM-related documents.

Separate statistics were calculated for the HTML and PDF pages in recognition that they may tend to be used for different purposes and so should be processed separately.

Noun Phrase Analysis

The noun phrase analysis extends the noun analysis to get more insights into the topics discussed in IWRM-related web pages by considering phrases containing nouns. These may give more specific topics details than nouns alone.

The noun phrase analysis was conducted similarly to the noun analysis above except that whole noun phrases were extracted from the part-of-speech tagged texts. Since phrases can contain other shorter phrases, maximal length noun phrases were first extracted and then all sub-phrases were extracted from the maximal phrases but single word "noun phrases" (i.e. individual nouns) were excluded since these would be included in the noun analysis.

Results and IWRM discussion

The data gathering method retrieved 41,381 IWRM-related HTML pages and 28,735 IWRM-related PDF documents. Comparative charts are presented in Figures 2-5 below of the top 30 links and URLs using various different aggregation methods, as described above and summarised in Table 1. The variety of different counting methods employed is undesirable for clarity of the results. Nevertheless, we believe that each selected method is superior to all alternatives and that an intuitive interpretation of the graphs and tables will be broadly correct in terms of the web data's reflection on the underlying issues.

Table 1. Counting methods used for pages and links

Data type	Measurement	Method
Domain/site	Size (pages)	The number of IWRM-related web pages with the domain/site name
STLD/TLD	Size (sites)	The number of sites containing IWRM-related web pages
Link target URL/domain/site	Frequency (unique inlinking sites)	The number of sites linking at least once to the target URL/domain/site from an IWRM-related web page
Link target STLD/TLD	Frequency (site pairs)	The number of unique pairs (site 1, site 2) such that at least one IWRM-related page in site 1 links to any page in site 2 with the link target URL matching the target STLD/TLD
Noun/noun phrase	Frequency (domains)	The number of web domains containing the identified noun/noun phrase

Link URLs compared to Source URLs

Figures 2 to 4 illustrate the most common origins of IWRM-related pages and targets of links in IWRM-related web pages. For example, in Figure 2, 11% of the IWRM source pages and 8% of the target URLs (counting target URLs after ignoring multiple link URLs originating from the same site) in the data set were from the .uk domain. The TLDs are standard international country codes (http://www.iana.org/cctld/cctld-whois.htm). The sites mentioned in the figures are not individually described, although some of them are clear from the domain names, others are mentioned in Table 2, and general trends are discussed in the text. The sites themselves can be visited online or via the Internet Archive (archive.org) to verify their identity in unknown cases.

International distribution

Broadly speaking, the distribution of source and target sites seen in Figure 2 is not surprising, being dominated by the countries and domains that have the largest representation on the web, predominantly the developed nations and with a bias towards English speaking nations (presumably exacerbated by the search phrases being English). This could well reflect web publishing strategies rather than underlying policy, which is not directly relevant to the aim of this study. Nevertheless, it is pertinent that there is not a significant trend against the dominance by developed nations, which was a possibility given that this is an issue where the potential beneficiaries are primarily developing nations. Also of interest are countries such as Denmark (.dk) and The Netherlands (.nl), which have a very high representation for target sites for this topic. In both cases this seems to reflect significant IWRM initiatives, although for The Netherlands the figures were slightly inflated by the hosting of an international site (Global Water Partnership Toolbox) with a Dutch domain name.

The com, org, and net sites reflect a combination of national and international organisations. The national organisations broadly reflect those of the TLDs, with the addition of some large US sites. The org sites included several UN or UN-sponsored initiatives as well as some national NGOs.

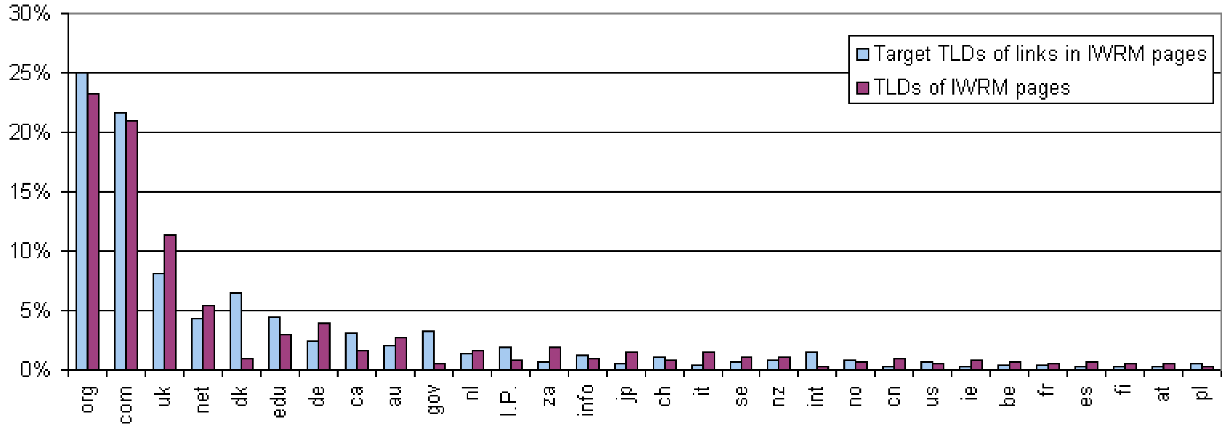

Fig. 2. The top IWRM TLDs (# of IWRM sites) and outlink target TLDs (# of unique linking site pairs).

Organisation type

Figure 3 seems to indicate a strong showing for non-profit organisations and companies, with a weaker but still significant participation of the education sector. This is misleading, however, because the majority of the com sites were non-profit organisations that use the com domain for convenience. The commercial sites that did exist were predominantly for academic and professional publishing, web site design, web site hosting, and commercially-hosted personal home pages. Hence, the underlying trend is for very strong NGO dominance, with some educational interest (edu, ac.uk) and a little commercial support, mainly for publishing. Although there are some commercial companies, such as environmental consultancies, that provide a non-publishing input into IWRM (Thelwall, Barlow et al., 2005), these are not evident on the web. This seems to indicate that these are probably rare, given that a company providing services for an international issue such as IWRM should naturally create a web site to attract clients.

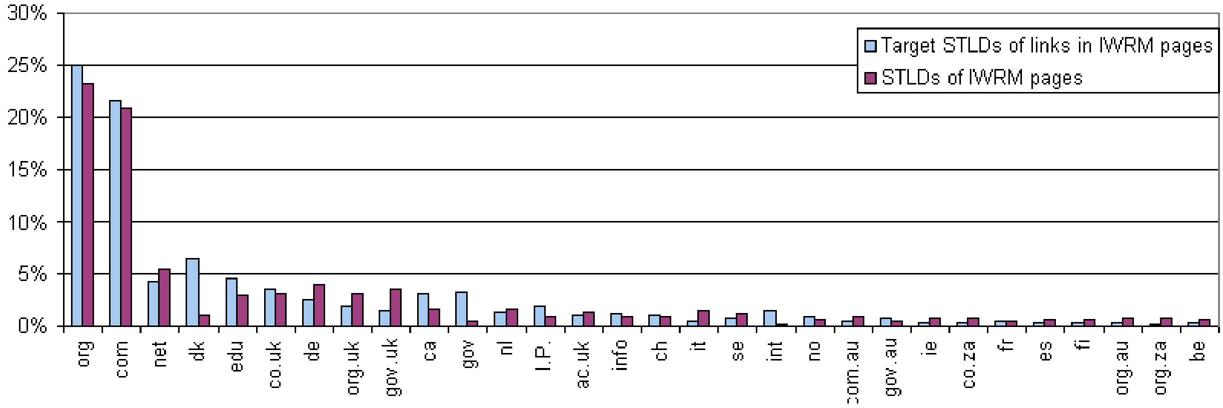

Fig. 3. The top IWRM STLDs (# of IWRM sites) and outlink target STLDs (# of unique linking site pairs).

Organisations represented

Figures 4 and 5 give information about the most common web sites and domains for IWRM-related pages and their links, and Table 2 gives the most highly targeted pages, together with some information about the organisation or initiative represented. The sites predominantly come from relevant major international organisations, national NGOs, and sites that provide web resources for NGOs. There are also web sites of the European Union and an international summit. The difference between publishing and linking is relevant in some cases, as some organisations publish a significant amount without attracting many links. This suggests that they have not (yet) had a significant impact on IWRM.

Three of the sites are unexpected: regional council sites from the UK and Ireland. Two of these are attempting to apply IWRM in their own areas, whereas the third is attempting to provide support for IWRM in an international context. None of the sites seem to have had a wider impact on IWRM, as evidenced by their lack of inlinks.

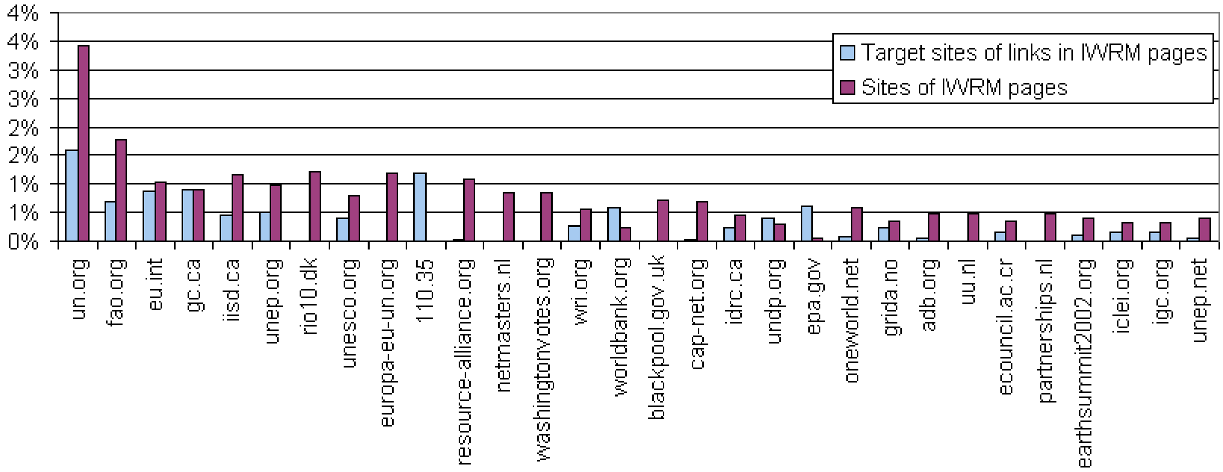

Fig. 4. The top IWRM sites (# of IWRM pages) and link target sites (# of unique inlinking sites).

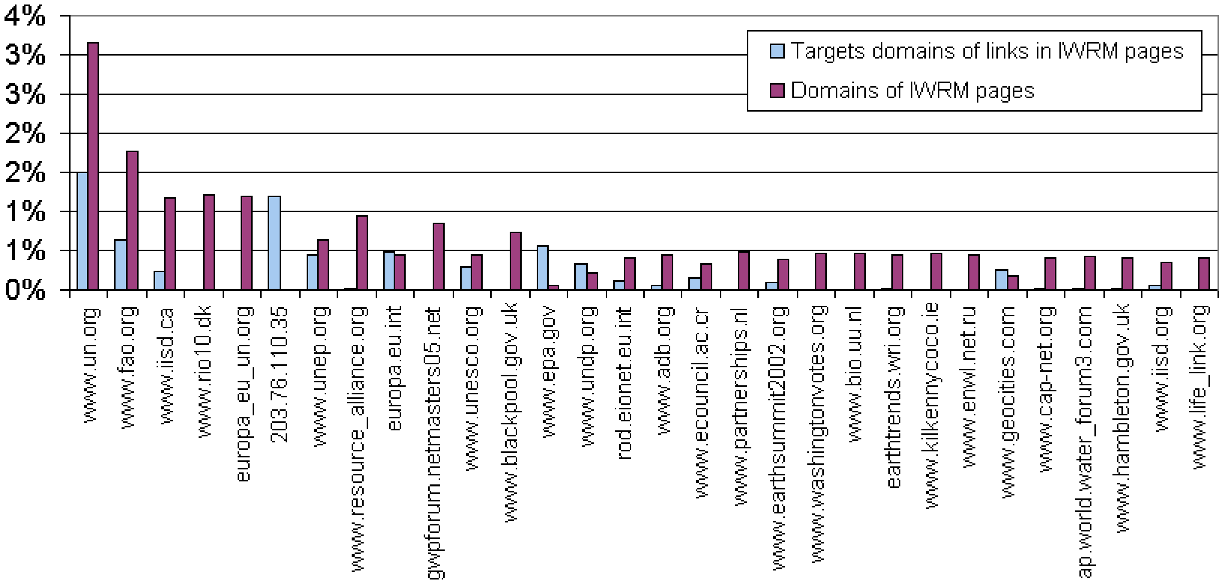

Fig. 5. The top IWRM domains (# of IWRM pages) and link target domains (# of unique inlinking sites).

Table 2. The top 30 inlinked pages.

# of unique inlinking sites	Link target URL	Site
349	http://www.unep.org	United Nations Environment Programme
344	http://www.johannesburgsummit.org	UN Johannesburg Summit 2002
213	http://www.undp.org	United Nations Development Programme
202	http://www.worldbank.org	The World Bank
188	http://www.iclei.org	International Council for Local Environmental Initiatives (ICLEI), a group of local government associations and organisations
184	http://www.biodiv.org	UNEP Convention on Biological Diversity
182	http://www.un.org/esa/sustdev	UN Division for Sustainable Development
177	http://www.fao.org	UN Food and Agriculture Organization
173	http://www.un.org	United Nations
172	http://www.iucn.org	IUCN -The World Conservation Union, international NGO
153	http://www.worldwatch.org	Worldwatch Institute, US environmental research NGO
152	http://www.epa.gov	US Environmental Protection Agency
140	http://www.wri.org	World Resources Institute, US environmental research NGO
134	http://www.environment-agency.gov.uk	UK government Environment Agency
133	http://www.panda.org	World Wildlife Fund
127	http://www.gwpforum.org	Global Water Partnership - UN-sponsored NGO
124	http://www.un.org/esa/sustdev/ agenda21.htm	The UN's Agenda 21 page.
118	http://www.greenpeace.org	Greenpeace, international environmental pressure group
114	http://www.ipcc.ch	Intergovernmental Panel on Climate Change (WMO and UNEP)
112	http://www.gefweb.org	Global Environment Facility, UNEP/UNDP/World Bank funding agency
111	http://www.unesco.org	United Nations Educational, Scientific and Cultural Organization
110	http://www.sustainable.doe.gov	US Department of Energy Center of Excellence for Sustainable Development
106	http://www.iucnrosa.org.zw	IUCN Regional Office for Southern Africa
106	http://www.eea.eu.int	European Environment Agency
105	http://www.earthsummit2002.org	Earth Summit 2002 (Johannesburg) web site
104	http://www.pwa-ltd.com	Philip Williams and Associates, environmental hydrology business
104	http://www.unfccc.de	United Nations Framework Convention on Climate Change
103	http://www.irchouse.demon.co.uk	Pisces Conservation Ltd - specialists in aquatic biology (UK based, international clients)
103	http://www.enservice.com	Environmental Services Inc. (works in New Mexico)
103	http://www.synectics.net	Data management service provider for the environmental industry

Site Interlinks

The site interlinking results are presented in the form of a network diagram of the top 50 most linked-to web sites (using the IWRM-related page links). In Figure 6, a link from site A to site B represents any IWRM-related page in site A containing a link to any page in site B (whether IWRM-related or not). Links to non-IWRM-related pages are included because links often take the form of targeting organisational home pages, allowing the user to navigate to pages of their choice. The diagram shows a

UN-dominated core of NGOs that heavily interlink, with a more peripheral set of NGOs and related organisations.

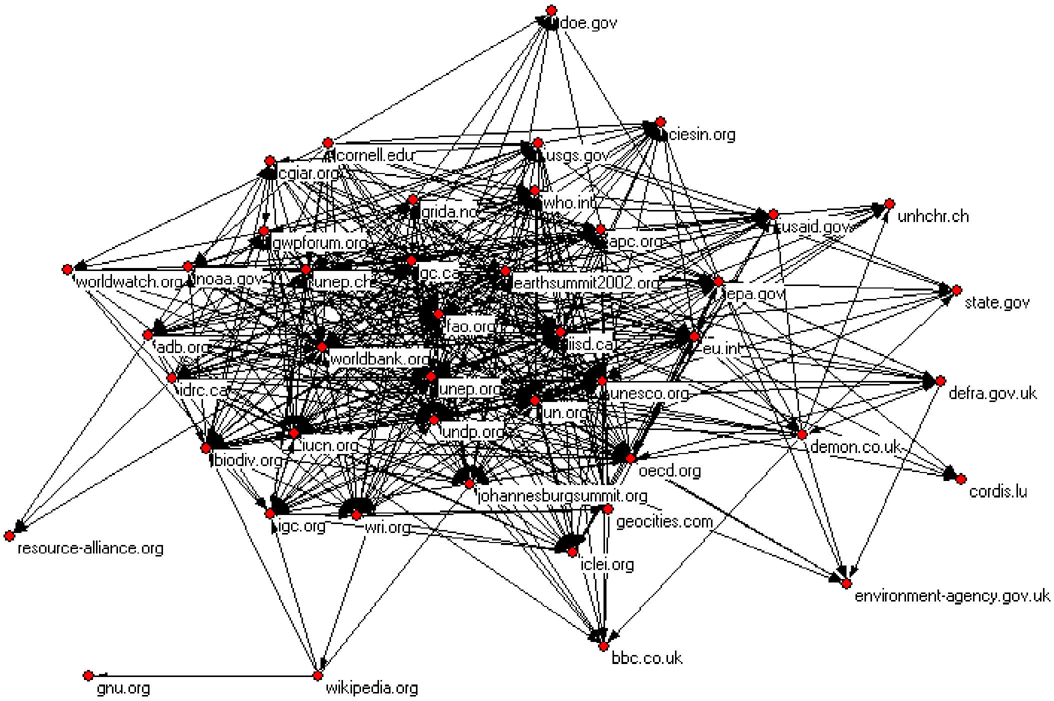

Fig. 6. Interlinking of the top 50 inlinked sites. An arrow represents at least one link in an IWRM-related page in the source site pointing to any page in the target site.

Nouns and noun phrases

For the text processing, the following statistics summarise the size of the data set.

- 8,380 domains contained at least one non-empty HTML page (i.e. containing at least one word).
- 5,468 domains contained at least one non-empty PDF document (i.e. containing at least one word).
- 1,322,733 noun phrases occurred in at least 2 separate domains in HTML pages.
- 2,193,168 noun phrases occurred in at least 2 separate domains in PDF documents.
- 191,417 nouns occurred in at least 2 separate domains in HTML pages.
- 277,401 nouns occurred in at least 2 separate domains in PDF documents.

Nouns

Figure 7 reports the relative frequencies of the most common words, after excluding all month names. The words probably reflect a combination of the issues that are most discussed and the organisations that are most involved in discussing them. Note that the percentages reported should be seen as minimums in most cases because the natural language processing algorithm does not always correctly classify words, and may misclassify important unusual text such as Agenda 21. Some exceptions are also present: “most” and “new” are not frequently occurring nouns in the corpus.

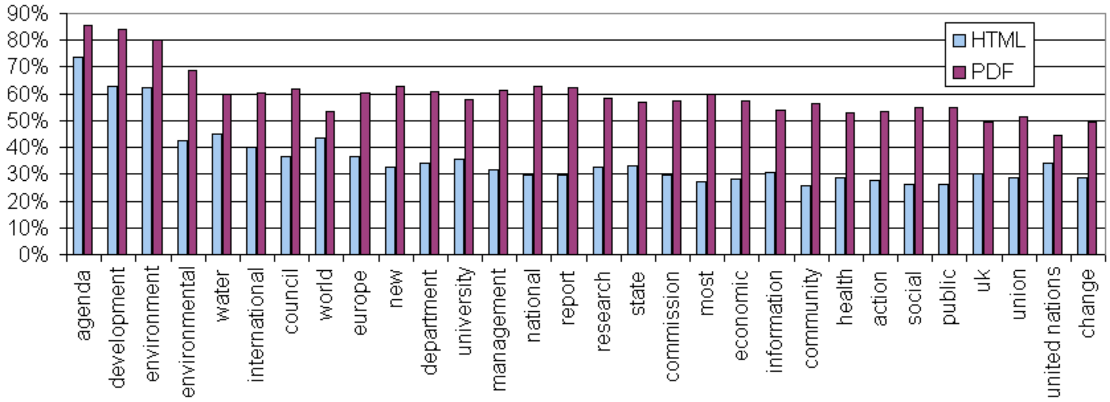

Fig. 7. Most common nouns, measured by the percentage of domains.

Noun Phrases

Figure 8 reports the most common multiple-word noun phrases. It is more informative than Figure 7 because the extra words give additional context. Nevertheless, the results are complimentary: for example Europe is found in Figure 7, and United States in Figure 8. The phrases include issues or concepts that appear to be particularly important but might not necessarily be immediately thought of in the context of IWRM, such as “future generations”, “private sector” and “economic growth”.

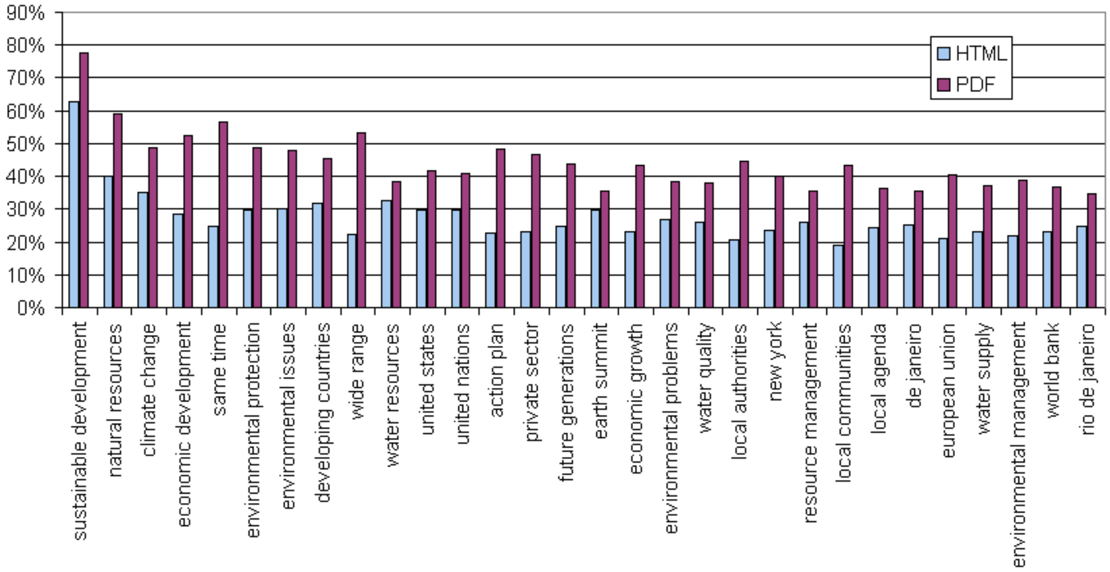

Fig. 8. Most common multiple-word noun phrases, measured by percentage of domains.

Discussion

Many of the details of data interpretation have been discussed in the methodology and results sections, but there are a few general points that are pertinent to a post-hoc evaluation and discussion of the methodology.

A practical problem with the approach used was the time taken for the natural language processing algorithms, which was approximately three weeks. In retrospect, it would have been better to take a sample of the domains, say just 1000, and process just these, because the overall results would have remained the same.

A language issue is that the initial searching for pages was conducted with English phrases. Although English is the standard language for international issues, it is likely that there would be pages in other languages that translate Agenda 21 and water and do not mention any of the phrases used here. These probably account for a small number of pages and hence this is probably not a major issue. In addition, the noun and noun phrase algorithms are English-specific, but again the expected low proportion of non-English pages about the issue means that this did probably not influence the results. Nevertheless, the issue of English bias is one that must be considered in any web issue analysis.

A human issue with the presentation of the results is that the method of counting is complex and so although the graphs are an attempt to present the information in an intuitive form, considerable effort would be needed to fully explain to stakeholders what all of the figures mean and why they were calculated in the form chosen. This is especially problematic because a range of different counting methods was employed, and such technical details will not be of interest to someone with a focus on IWRM itself. This is a fundamental problem for the general methodology of this paper: it remains to be seen whether figures presented in such a form will be acceptable to users of the information, or whether it is seen as too complex or obscure. Nevertheless, the noun and noun phrase statistics are relatively straightforward: counting the number of domains containing each one, and these may prove to be the most useful statistics from this kind of analysis in the long term.

Conclusion

In general terms, the page and link statistics presented above suggest the domination of IWRM by NGOs, and by developed nations with particular contributions from Denmark and the Netherlands. The noun and noun phrase results emphasised the importance of sustainable development as the key issue for IWRM, being discussed in about 70% of cases. A more fine-grained analysis of the results may be useful for a stakeholder to identify the types of organisations represented, and the most prolific and linked-to organisations. It can also be used to fill in gaps in knowledge for experienced participants, and as a useful quantitative confirmation of existing beliefs. One example of information that seems to be largely hidden is that a few local governments in developed nations have adopted IWRM initiatives (many IWRM-related pages, but few links from other IWRM-related pages). This web issue analysis information could also be used to help initiate IWRM newcomers into the structure of the field. This is a similar claim to that made by Author Cocitation Analysis (White & McCain, 1998). In the wider context of using information gathered to produce this paper, the graphs and tables include only 30 entries for reasons of space, but when presenting the results to stakeholders, the full spreadsheets can be given to them in electronic form so that they can be browsed in more depth to identify information that is new or unexpected.

It would be interesting to apply similar web issue analyses over time to see how the results change, as has been achieved for database-based issue analyses (Lancaster & Lee, 1985). It would be particularly useful in cases where an event occurred, such as a new policy decision being made, and an analysis was conducted before and after, showing the event’s impact, at least on the web.

The results presented here show that web issue analysis can present information about the organisations and issues relevant to a given theme through a mainly automated processing of data from the web. Whilst the interpretation of the results needs to be conducted carefully, for example because of misleading site names and the over-representation of web publishing countries, it has the potential to be a fast and independent method for exploring issues via the web.

Acknowledgements

We thank the referees for insightful and valuable comments and Google for allowing researchers to submit automated queries via its API. This research was supported by the Netherlands Institute for Scientific Research (NWO - Shifts in Governance Programme Award 450-04-135) and a British Council/NWO Partnership in Science Programme Award PPS 804.

References

Alfarra, A. (2004). *Modelling water resource management in Lake Naivasha.* International Institute for Geo-information Science and Earth Observation.

Bar-Ilan, J. (2000). The Web as an information source on Informetrics? A content analysis. *Journal of American Society for Information Science, 51*(5), 432-443.

Bar-Ilan, J. (2004a). A microscopic link analysis of academic institutions within a country - the case of Israel. *Scientometrics, 59*(3), 391-403.

Bar-Ilan, J. (2004b). The use of Web search engines in information science research. *Annual Review of Information Science and Technology, 38*, 231-288.

Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management, 41*(3), 973-986.

Bar-Ilan, J., & Peritz, B. C. (1999). The life span of a specific topic on the Web. The case of "informetrics": A quantitative analysis. *Scientometrics, 46*(3), 371-382.

Bar-Ilan, J., & Peritz, B. C. (2004). Evolution, continuity, and disappearance of documents on a specific topic on the Web: A longitudinal study of 'informetrics'. *Journal of the American Society for Information Science and Technology, 55*(11), 980 - 990.

Bharat, K., & Henzinger, M. (1998). Improved algorithms for topic distillation in a hypertext environment. *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 104-111.

Björneborn, L. (2004). *Small-world link structures across an academic web space - a library and information science approach.* Royal School of Library and Information Science, Copenhagen, Denmark.

Blatter, J., & Ingram, H. M. (Eds.). (2001). *Reflections on water: New approaches to transboundary conflicts and cooperation*. Cambridge, MA: MIT Press.

Bonell, M. (2004). How do we move from ideas to action? The role of the HELP Programme. *International Journal of Water Resources Development, 20*(3), 283-296.

Brill, E. (1992). A simple rule-based part of speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, 152-155.

Calder, I. (1999). *The blue revolution: land use & integrated water resources management*. London: Earthscan.

Chakrabarti, S., Joshi, M. M., Punera, K., & Pennock, D. M. (2002). *The structure of broad topics on the Web*, from http://www2002.org/CDROM/refereed/338

Chu, H. (2005). Taxonomy of inlinked web entities: What does it imply for Webometric research? *Library & Information Science Research, 27*(1), 8-27.

Cimiano, P., & Staab, S. (2004). Learning by Googling. *SIGKDD Explorations, 6*(2), 24-33.

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). Self-organization and identification of Web communities. *IEEE Computer, 35*, 66-71.

Foot, K. A., Schneider, S. M., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 US electoral web sphere. *Journal of Computer Mediated Communication, 8*(4), http://www.ascusc.org/jcmc/vol8/issue4/foot.html.

Garrido, M., & Halavais, A. (2003). Mapping networks of support for the Zapatista movement: Applying Social Network Analysis to study contemporary social movements. In M. McCaughey & M. Ayers (Eds.), *Cyberactivism: Online activism in theory and practice* (pp. 165-184). London: Routledge.

Heimeriks, G., Hörlesberger, M., & van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics, 58*(2), 391-413.

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

Introna, L., & Nissenbaum, H. (2000). Shaping the web: Why the politics of search engines matters. *The Information Society, 16*(3), 1-17.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM,, 46*(5), 604-632.

Koehler, W. (2004). A longitudinal study of Web pages continued: a report after six years. *Information Research, 9*(2), 174.

Lancaster, F. W., & Lee, J. l. (1985). Bibliometric techniques applied to issues management - a case-study. *Journal of the American Society for Information Science, 36*(6), 389-397.

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

Levesque, S. (2001). The Yellowstone to Yukon conservation initiative. In J. Blatter & H. M. Ingram (Eds.), *Reflections on water: New approaches to transboundary conflicts and cooperation*. Cambridge, MA: MIT Press.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy, 18*, 209-223.

Leydesdorff, L., & Hellsten, I. (2005, to appear). Metaphors and diaphors in science communication: Mapping the case of 'stem-cell research'. *Science Communication*, http://www.leydesdorff.net/stemcells.pdf.

Maganga, F., Kiwasila, H., Juma, I., & Butterworth, J. (2004). Implications of customary norms and laws for implementing IWRM: findings from Pangani and Rufiji basins, Tanzania. *Physics and Chemistry of the Earth, 29*(15-18), 1335-1342.

Mair, C. (2003). *Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora.* Paper presented at the ICAME conference, Guernsey.

Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines - fluctuations in document accessibility. *Journal of Documentation, 57*(5), 623-651.

Meyer, C., Grabowski, R., Han, H.-Y., Mantzouranis, K., & Moses, S. (2003). The world wide web as linguistic corpus. *Language and Computers, 46*(1), 241-254.

Mitkov, R. (2003). *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press.

Musgrove, P. B., Binns, R., Page-Kennedy, T., & Thelwall, M. (2003). A method for identifying clusters in sets of interlinking Web spaces. *Scientometrics, 58*(3), 657-672.

Newson, M. (1999). *Land, water and development: Sustainable management of river basin systems* (2 ed.). London: Routledge.

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections, 25*(1), 49-61.

Polanco, X., Boudourides, M. A., Besagni, D., & Roche, I. (2001). Clustering and mapping Web sites for displaying implicit associations and visualising networks: University of Patras.

Price, E., & Thelwall, M. (2005). The clustering power of low frequency words in academic webs. *Journal of the American Society for Information Science and Technology, 56*(8), 883-888.

Rogers, R. (2002). Operating issue networks on the Web. *Science as Culture, 11*(2), 191-214.

Smith, A. G. (1999). A tale of two web spaces; comparing sites using Web Impact Factors. *Journal of Documentation, 55*(5), 577-592.

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

Thelwall, M. (2003). A layered approach for investigating the topological structure of communities in the Web. *Journal of Documentation, 59*(4), 410-429.

Thelwall, M. (2004a). *Link analysis: An information science approach*. San Diego: Academic Press.

Thelwall, M. (2004b). Methods for reporting on the targets of links from national systems of university Web sites. *Information Processing and Management, 40*(1), 125-144.

Thelwall, M. (2004c). Vocabulary Spectral Analysis as an exploratory tool for Scientific Web Intelligence. In E. Banissi (Ed.), *Information Visualization (IV04)* (pp. 501-506). Los Alamitos, CA: IEEE.

Thelwall, M. (2005). Text characteristics of English language university web sites. *Journal of the American Society for Information Science and Technology, 56*(6), 609-619.

Thelwall, M., & Aguillo, I. F. (2003). La salud de las Web universitarias españolas. *Revista Española de Documentación Científica, 26*(3), 291-305.

Thelwall, M., Barlow, A., & Vann, K. (2005). The limits of web-based empowerment: Integrated Water Resource Management case studies. *First Monday, 10*(4), Retrieved April 20, 2005 from: http://www.firstmonday.org/issues/issue2010_2004/thelwall/.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology, 39*, 81-135.

Thomas, J. S., & Durham, B. (2003). Integrated water resource management: Looking at the whole picture. *Desalination, 156*(1-3), 21-28.

United_Nations. (1992). *Agenda 21: The United Nations programme of action from Rio*. New York: United Nations.

Vaughan, L., & Hysen, K. (2002). Relationship between links to journal Web sites and impact factors. *ASLIB Proceedings, 54*(6), 356-361.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management, 40*(4), 693-707.

Vaughan, L., & Wu, G. (2004). Links to commercial websites as a source of business information. *Scientometrics, 60*(3), 487-496.

Weare, C., & Lin, W. Y. (2000). Content analysis of the World Wide Web-Opportunities and challenges. *Social Science Computer Review, 18*(3), 272-292.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327-355.

Wormell, I. (2000). Critical aspects of the Danish Welfare State - as revealed by issue tracking. *Scientometrics, 48*(2), 237-250.

Wright, E. (2003). The re-design of an integrated water and pollution management programme using the systems-ware model of the log frame. *Physics and Chemistry of the Earth, 28*(20-27), 973-984.