Information retrieval from the Internet for academic use and copyright issues

J. Reyneveld and M. Weideman
e-Innovation Academy, Cape Peninsula University of Technology

Abstract

This research study discusses the reasons for using search engines vs. academic search engines for information retrieval, and the importance of copyright. The Internet was used to gather information for the study, in which the relevance of the different search engines for different end users is indicated. This research took into account that every searcher may have a different need, but it focuses more on the need for academics to identify good sources of information.

The result shows that academic search engines are better suited if in need of information of an academic nature, be it sciences or history. It also proved that an academic search could present more accurate information, as previous researchers have investigated one's topic, or something similar, on previous occasions. The conclusion was drawn that to base one's investigation for information solely on popular search engines could leave one with information that is not viable because anybody could have placed it on the Internet. With academic databases one can be assured to a great extent that the information retrieved comes from a credible source.

Keywords

Information retrieval, copyright, search engines, citation

INTRODUCTION

This research addresses an information retrieval process using the Internet as the primary source and some of the factors that academics need to be aware of. These include copyright, evaluating the usefulness of websites, tips on identifying good or bad sites and the best source for information retrieval.

A wide range of opportunities exists for finding information on the Internet, with online focus groups and online questionnaires to only mention a few. It is safe to say that the Internet opens up new possibilities for research.

With all these sources available on the net, researchers need to be cautious of certain factors to identify a good source.

This article attempts to discuss the reasons for using search engines vs. academic search engines or directories as means of information retrieval, and the need to take copyright issues into consideration when making use of information from the Internet.

LITERATURE SURVEY

Search engines

A search engine is an information retrieval system designed to help find information stored on a computer system.

Page and Brin (1999) had the following to say about search engines. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them.

Academic search engines

Academic search engines are also known as academic databases, stored articles, books, journals, etc.

Most academics rely on specialist databases to access details of past research. The databases draw together details of scholarly publications from a wide range of sources including academic publishers, journals, archives and sometimes books, and enable one to search a large body of the scholarly literature in one go (Place et al., 2006).

Anyone can publish information on the Internet. The problem/challenge is evaluating the large amount of Internet resources for its usefulness and content.

The Internet can be used for a number of reasons: for literature reviews, data collection, data analysis and also for disseminating information.

Arnzen (1997) discusses the pros and cons of using Internet sources in his article.

Many students, teachers, writers, and research scholars see the Internet as a vast virtual library full of books, essays, and articles. But there are two problems with considering the Internet as a virtual library. For one thing, the documents on the Internet, unlike the books in the library, are difficult to catalogue. The Internet's holdings change every minute of the day, taxing even the best search engines. Another problem with the virtual library metaphor is that many Internet 'books' can disappear off the shelf with the touch of a webmaster's

delete key, leaving the student in a quandary if he or she wants to borrow the book again – or worse, leading to charges of poor ethics if a professor can't verify a source cited in a research paper (Arnzen, 1997).

Many Internet-based papers are electronic versions of peer-reviewed journal papers. As such, they are likely to be trustworthy, especially if they are hosted on the journal publisher's website, or if the author provides a full reference to the paper-based version (Staff, 2001).

Accuracy of the information gathered can only be truly judged on the basis of some prior knowledge. If preliminary reading has been completed on a topic, the reader would be in the best position to judge whether a site contains enough accurate information. Judgement on whether information is adequate, as well as whether all sides of an argument are shown can then be evaluated (Cameron, 2001).

Cameron (2001) claims that the following points should be considered when viewing a site for the first time:

- Is it clear who wrote the content, and how to contact them? Do they give information about their background and qualifications?
- Is there an editor, and if so, who is it? If there is no author or editor information, why is the site owner so keen to remain anonymous?
- Is the aim of the site stated clearly and does the content match that aim?
- Does the spelling, grammar and composition conform to accepted standards? If they cannot even get that right, can you trust their ideas?
- Do the external links lead to sites you know to be trustworthy? Do any trustworthy sites link to this one?
- Is the information given verifiable – does the writer quote sources?

These factors might have been listed some time ago, but still predominantly hold true today.

Guidelines for checking your sources

The researcher should check for statements of authorship left by the creator. Another good guideline is to do a quick Internet search on the author's name if they are not well known. It could be beneficial to check whether the author has any other publications on a relevant bibliographic database. When quoting information from the website of an organisation the researcher should make sure that the information is coming from a reputable body (Place et al., 2006).

Cull (2006) also lists his general criteria to consider when evaluating a whole publication:

- **Accuracy** Whether the facts are correct?
- **Authority** Is the author listed? What are the author's qualifications?
- **Objectivity** What is the purpose of the site? Is there any sign of preconception?
- **Currency** Is a date given? When it was last updated?
- **Coverage** When evaluating the coverage of the topic (as explained under criteria #3) remember to evaluate individual sites separately.

Worley (2006) discussed the topic of limitations of search engines, and stated that search

engines are one of the least useful ways of finding large quantities of relevant legal information, and further stated that the search engines are only valuable if all other Internet avenues have failed. The author comments that the Internet is growing faster than search engines can index it, and that search engines cover no more than 17% of the entire Internet.

The parts of the Internet containing the most useful information are hidden from search engines, whether behind passwords, registration forms or the fact the search engines cannot search PDF files. Search engines take months to capture a new site as well as indexing and adding it to the database, not to mention that sometimes search engines are programmed to collect from the most popular sites only (Worley, 2006).

The conclusion is that search engines are better at collecting news, current affairs and leisure sites rather than technical or academic sites (Worley, 2006).

Copyright

Copyright is a set of exclusive rights that regulate the use of a particular expression of an idea or information. In general, it is literally 'the rights to copy' an original creation. In most cases, these rights are of limited duration. The symbol for copyright is '©'.

The following words are closely linked to copyright:

- Plagiarism,
- Copyright, and
- Referencing and citation.

By not acknowledging sources the researcher could be accused of plagiarism or of being in breach of the copyright law (Place et al., 2006).

Besek et al. (2003) stated that copyright exists in any original work of authorship fixed in a tangible medium. That medium can be almost anything, including paper, computer disk, clay, canvas, etc. Copyright lasts for the life of the author and 70 years thereafter. For a work to be original, it must meet two qualifications:

- It cannot be copied from another work, and
- It must exhibit at least a small amount of creativity.

Academics usually publish their research in formal publications such as journal papers and articles or reports. These follow formal procedures designed to quality-assure the work. Plagiarism should be avoided by always acknowledging your sources and by citing and referencing them correctly (Place et al., 2006).

Weideman (2005) lists the proper way to reference sources for academic use.

Peer review/referencing

Peer review is what characterises academic research. If a publication is peer-reviewed it means it has been read, checked and authenticated (reviewed) by independent, third party academics (peers). Peer review has been the quality-control system of academic publishing for hundreds of years.

Scholarly journals

Peer-reviewed articles are often collated into scholarly journals, which are usually published by academic publishing houses, professional societies or a university press.

Electronic journals

A university library may have shelves full of journals, but nowadays many are also available in electronic form over the Internet.

Bibliographic databases

Most academics rely on specialist databases to access details of past research. The databases draw together details of scholarly publications from a wide range of sources including academic publishers, journals, archives and sometimes books, and so enable the researcher to search a large body of the scholarly literature in one go.

Library websites

The library website for a university or college will be an important source of information, as it guides you to the key electronic journals, bibliographic databases and archives.

RESEARCH METHOD

Bond (2004) completed a research project investigating how people approach the task of finding information on the Internet. The study was conducted on 18 staff members of a nursing department who were to find a specific document pertaining to the Department of Health. The participants had a range of skills and experience in using the Internet, and ranged between ages 19 and 63.

Only three of the participants eventually found the information, but the main conclusion was that there were three ways that the respondents had first tried to find the information:

- Going directly to the Department of Health website by entering the web address;
- Some used the default search option available within the browser to try to find the Department of Health website; and
- Some selected their own preferred search engine to find either the Department of Health or the document itself.

The conclusion was that the first choice of the majority of participants, when faced with the task of finding a specific document produced by a well known organisation, was to go to (or try to go to) the organisation's site, and to locate the document from there. The most successful way of finding the document, however, was by the efficient use of a search engine.

The following example illustrates the use of a search engine for different end results: Which of the sites in Figure 1 would better suit your need for articles to support your dissertation research? The answer: *d) Journal home: Nature*. Opening the link would show that *Nature* is an international peer-reviewed journal of science and this would be an appropriate starting point to begin research.

Which of the above sites would be better suited for a 12-year-old's homework? The

answer: *c) The EPA Global Warming Kids Page*. This site is designed specifically as a teaching and learning tool aimed at providing school children with a basic overview of the issues of global warming; this would be an ideal place to look for information (Place et al., 2006). Place et al. (2006) draws a good picture of investigating a suspect source.

URL

Figure 2, below, shows the URL for the website in Figure 3. The URL in Figure 2 suggests a non-governmental, non-profit-making organisation, which seems appropriate

Contact Address

The postal address shows that the organisation is based in Geneva in Switzerland, and a full address is listed, which again seems appropriate.

FIGURE 1: Search results for the topic Global Warming

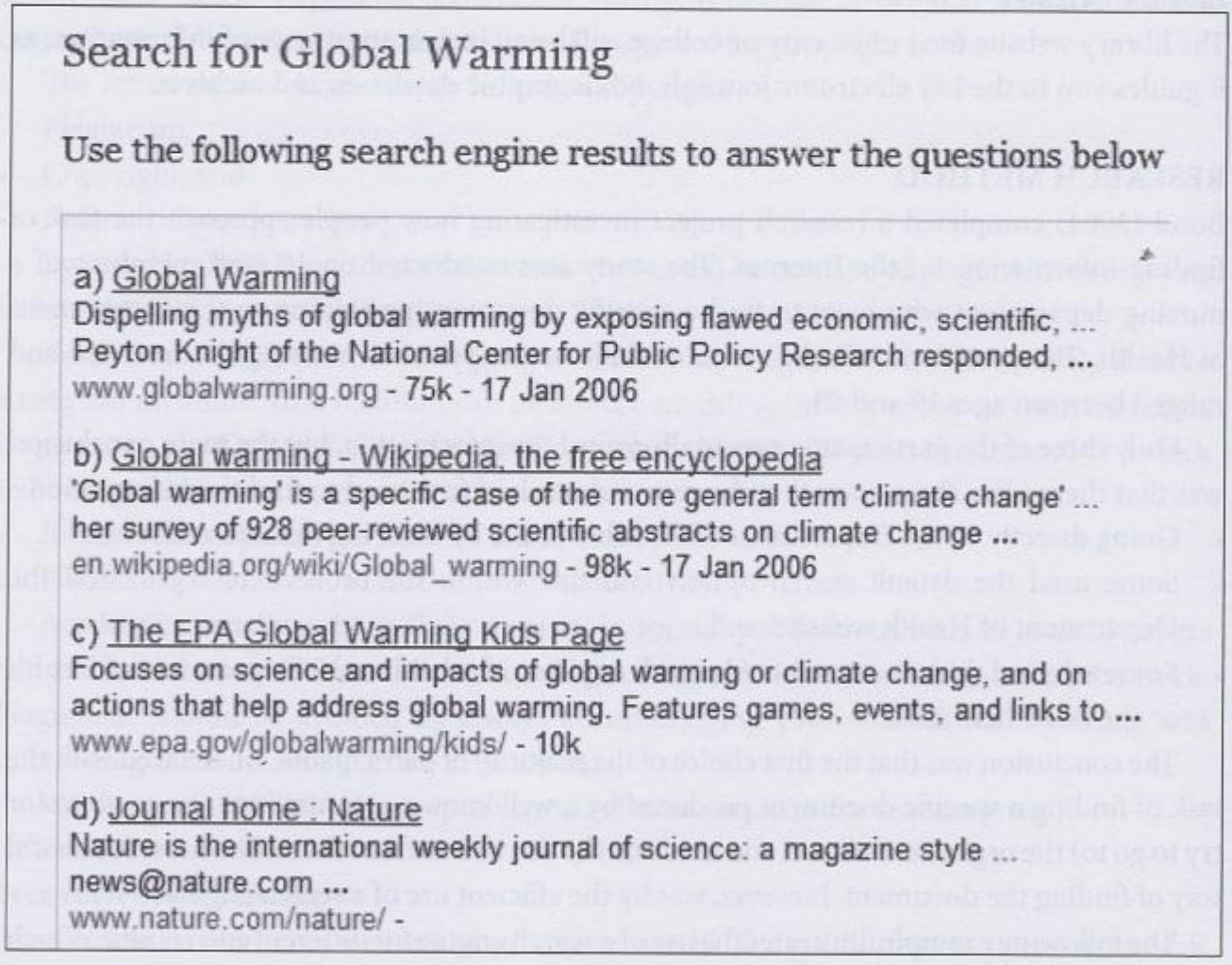


FIGURE 2: URL

Investigate the content

The link describing the WTO itself has a press release dated 20 May 2002 announcing that the WTO will be disbanding – this could be an indication that something is wrong and it might be beneficial to investigate this.

Checking the registration of the URL

Check to see who has registered the domain name of a site by using Register.com (http://www.register.com, 2007). This reveals that the gatt.org is registered to an individual in the

FIGURE 3: Example of WTO website

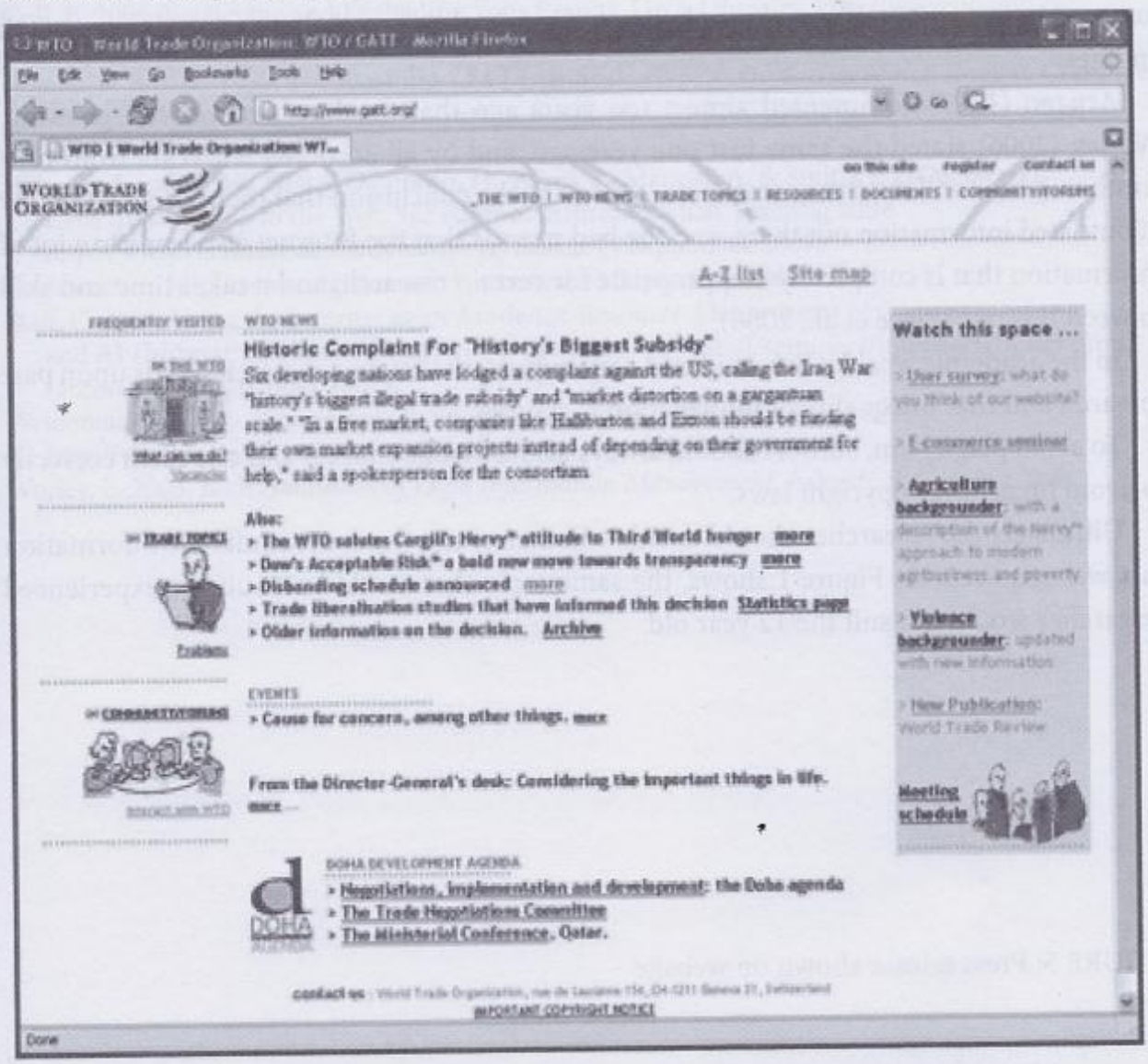
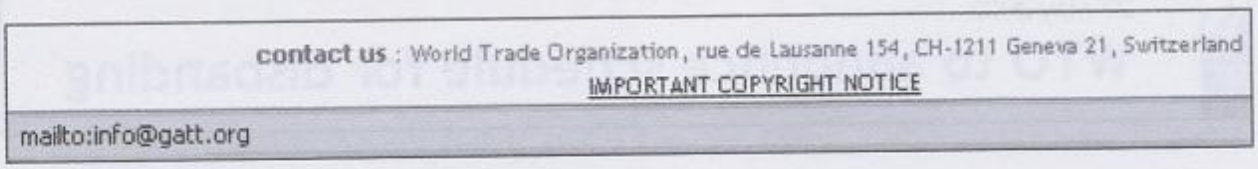


FIGURE 4: Show copyright and contact information



States, which seems unlikely given that the WTO is a large international organisation (Place et al., 2006).

Place et al. (2006) concluded that this is actually quite a sophisticated hoax site set up by a group of anti-WTO campaigners. It is designed to look almost identical to the real site and even the URL is using the name of the predecessor organisation of the WTO – the General Agreement on Tariffs and Trade (Gatt), which makes it seem more plausible.

CONCLUSIONS AND RECOMMENDATIONS

Having searched the Internet using both conventional search engines, as well as academic search engines, it is clear that there is a feeling that the best source for information retrieval for academic use would be through the use of academic websites or academic search engines.

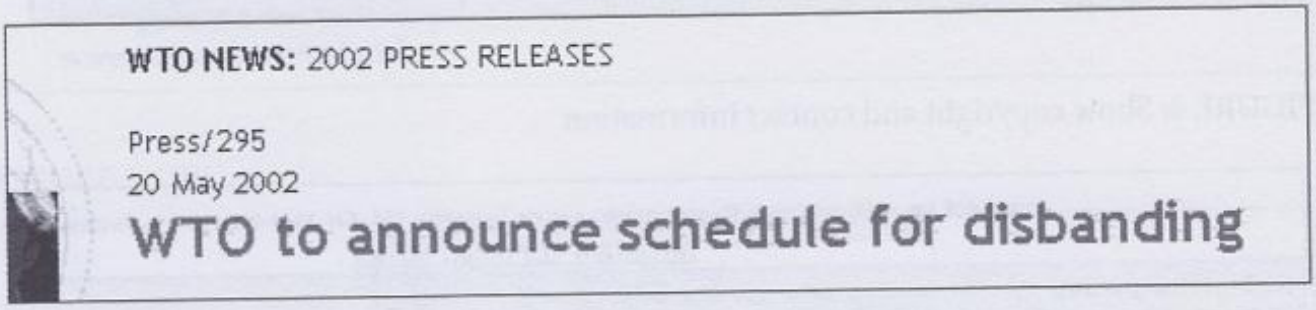
Arnzen (1997) commented almost ten years ago that the Internet was fast growing. Worley (2006) stated the same just one year ago, and by all accounts ten years from now another author will state the same. This leads us to the conclusion that there could be plenty of outdated information out there, and the bad news is that the Internet also leads to a lot of information that is completely inappropriate for certain research, and it takes time and skill to weed this out (Place et al., 2006).

In the academic world it is considered very important that new research builds upon past research and that the quality of information is assured.

To avoid plagiarism, authors should always acknowledge sources and cite them correctly to avoid breaching copyright laws.

Ultimately the researcher should decide whether or not the source suits the information that was required. As Figure 1 shows, the same topic that would best suit the experienced researcher would not suit the 12 year old.

FIGURE 5: Press release shown on website



REFERENCES

Arnzen, M.A. 1997. 'Documenting Internet sources presents some thorny problems'. *Cyber-citations.* http://www.rohan.sdsu.edu:80/faculty/krentler/geninfo/cybercitations.htm (accessed 17 September 2007).

Besek, J.M. 2003. Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessment. Council on Library and Information Resources Washington, D.C. and Library of Congress. http://www.clir.org/pubs/reports/pub112/pub112.pdf (accessed 15 September 2007).

Bond, C.S. 2004. Web users information retrieval methods and skills. Emerald. www.emeraldinsight.com (accessed 19 September 2007).

Cameron, S. 2001. Evaluating Internet Sites for Academic Use. The Higher Education Academy. http://www.hca.heacademy.ac.uk/resources/Briefing_Papers/index.php (accessed 10 September 2007).

Cull, B. 2006. Retrieving and Evaluating Your Results. UNB Libraries. http://www.lib.unb.ca/instruction/InfoSearch_Retrieving.html (accessed 14 October 2007).

Page, L. & Brin, S. 1999. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Science Department, Stanford University, Stanford, CA 94305. http://infolab.stanford.edu/~backrub/google.html

Place, E., Kendall, M., Hiom, D., Booth, H., Ayres, P., Manuel, A. & Smith, P. 2006. *Internet Detective: Wise up to the Web*. 3rd edition, Institute Virtual Training Suite. http://www.vts.intute.ac.uk/detective (accessed 11 September 2007).

Register.com. http://www.register.com/ (accessed 13 September 2007).

Staff, C. 2001. Using the Internet as an Academic Resource. Department of Computer Science and AI University of Malta. http://www.cs.um.edu.mt/~cstaff/seminars/internetSeminar.html. (accessed 10 September 2007).

Weideman, M. 2005. Bibliograpic Referencing. Academic Research Resources. http://www.mwe.co.za/itresresources.htm#2. (accessed 15 July 2007).

Worley, L. 2006. *Biall Handbook of Legal Information Management.* Ashgate Publishing, Ltd.