Issues in Informing Science and Information Technology Volume 3, 2006

A Principled Methodology for Information Retrieval on the Web

Martin Eayrs
University of Salford, Manchester, UK

m.eayrs@salford.ac.uk

Abstract

The World Wide Web provides a wealth of information - indeed, perhaps more than can comfortably be processed. But how does all that Web content get there? And how can users assess the accuracy and authenticity of what they find? This paper will look at some of the problems of using the Internet as a resource and suggest criteria both for researching and for systematic and critical evaluation of what users find there.

Keywords: critical thinking, www, unprincipled surfing, principled surfing, accuracy, authority, objectivity, currency, coverage

Introduction

"It's not what you don't know that'll hurt you - it's what you do know, that isn't so"

(Arase, 2003)

Is there a danger that over reliance on the Internet as a source of information for academic purposes might result in a degradation of the integrity of data acquired? Is the anarchical and pioneering nature of much of the Internet compatible with the provision of reliable and authoritative resource material? This paper considers these questions, and suggests that balanced and critical consideration of the origin and purpose of web texts may be beneficial to students, and may constitute a useful study skill for teachers to provide students with.

Content

The Importance of Critical Thinking

The ability to think critically is what allows a reader to see beyond the literal meaning of discourse and give balanced, reasoned and well-grounded responses. The same evaluative criteria should be brought to bear on the selection of appropriate resources for teaching or research purposes.

Critical thinking is, in short, the ability to perceive and respond appropriately to a great deal that is - and is not - immediately apparent to the casual reader or viewer. Critical thinking takes the learner beyond the literal, surface aspects of any given piece of material into the motivation and purpose of the writer.

This paper will not attempt to define ‘critical thinking’, but does suggest that certain principles – in particular, the ability to evaluate - can be applied to the evaluation of material found everywhere, and particularly, for the purpose of this paper, on the WWW.

Reading in the Offline World

When people pick up something to read they already have certain expectations. To use examples from UK newspapers, one tends to approach The Sunday Telegraph rather differently from The News of the World. Previous reading experience has led readers to identify each as a separate ‘type’ and the reader reacts to each on the basis of this experience.

‘Periodicals’, to take a category at random, can be divided into four categories:

- Scholarly - ‘concerned with academic study, especially research’
- Substantive News / General Interest - ‘having a solid base, being substantial’
- Popular - ‘fit for, or reflecting the taste and intelligence of, the people at large’
- Sensational - ‘arousing or intending to arouse strong curiosity, interest or reaction’

(Merriam-Webster's Online Dictionary, 2006)

It is experience and familiarity with these broad genres that enable readers to assign a publication to one of these categories, and although there will be borderline cases readers on the whole feel confident about so assigning them. What is interesting, however, is how they are able to do this. This paper suggests that critical thinking plays an important part.

This ‘grading’ system works well for printed periodicals, where on opening the publication the reader’s expectations are largely met. However, when following the results of a search engine or a link on a web page it is not always clear what lies in store, and all may not always be what it seems. For example, it is uncommon to find advertising placed alongside the text of printed academic articles, and flashing sprites or animated gifs that waltz across the page to the accompaniment of midi or other sound files, are not a feature of academic journals. And whereas academic journals tend to be produced by distinguished scholars in the ivory towers of academe, all the world and his brother seems to be producing web-pages, some built and maintained by children as young as seven or eight.

Put at its simplest, when readers open a book or journal they generally have a clear idea where their reading will take them. When the same reader clicks on a hypertext link on the World Wide Web they can't really be sure exactly where they are going, and even when they get there they can’t always be sure exactly where they are. But there are clues, as shall be seen.

The Dangers of Unprincipled Sourcing and Referencing

As has been stated, academics all have to learn how to distinguish 'scholarly' journals from other periodicals. In fact these might be divided these into such areas as: 'academically respectable', 'professionally solid', 'popular' and 'sensationalist' - as was done above - but how does a reader do this? What criteria is brought to bear?

When evaluating traditional printed material there are certain points one would do well bear in mind. These are well known and generally taught as part of an academic skills programme. They are presented below in the form of questions one might wish to ask, drawing on Susan Beck (1997).

Accuracy

- Is the information reliable?
- Has the information been checked for errors of fact? By whom?
- Is there an editor who assumes authority for this?
- Are there people entrusted with the specific duty of checking all the facts?
- Is it clear who these people are?

Authority

- Is the author really ‘qualified’ to write on this subject?
- Is the author a recognised authority in this field?
- Is the publisher a ‘reputable’ publisher?
- Is the book published privately, perhaps even by the author himself?

Objectivity

- Is bias kept to a minimum?
- Is the information trying to influence the opinion of the audience?
- Is theory supported by data or coherent argument?
- Is unsupported theory built on to produce subsequent argument?

Currency

- Is the content of the work up-to-date?
- Is the date of publication clearly marked?
- Can one be satisfied that the given date is genuine?

Coverage

- Is the range of topics included in the work adequate?
- Is the range of topics included in the work relevant?
- Is the range of topics included in the work explored in reasonable depth?

Criteria for Web Evaluation

These criteria, remember, are used for evaluating printed material. Now these same points will be applied to publishing on the World Wide Web.

Accuracy

The truth – for good or bad - is that anyone with the right software and who knows how to type and navigate can publish on the Web. Statistically, few Web resources are verified by editors or fact checkers and there are as yet no internationally accepted Web Standards to ensure the accuracy of information displayed. In the worst of cases, there is nothing whatsoever to prevent the malicious posting of totally false (or, what is far worse, partially false) information.

Authority

It is usually difficult and often impossible to determine real authorship of material posted on the Web. In cases where an author’s name is listed his/her qualifications are usually not, and responsibility for accuracy or truthfulness of content is very rarely stated. The fact that an article is attributed to, say, Noam Chomsky, is no guarantee that Chomsky himself wrote it or that it even reflects his views.

Objectivity

It is not claimed here that writers should not attempt to be persuasive their writing, but there is an academic tradition of presenting alternative arguments and interpretations, and in most writing the literature revue will refer widely to the field. In. But in many cases the aims of persons or groups who post material on the web are often unclear. The Web has a tendency to act as a "virtual soapbox" where anybody can get up and say what they want - indeed this is one of its strengths - but it certainly attracts a fair share of eccentrics along with the altruistically motivated. Which of these is which is not always clear.

Currency

Any active Internet user is aware of the number of dead pages littering the Web; sites, for example, saying 'last modified August 1995'. 1995 is in fact an interesting date, as it was about then that postgraduate students started to have university web pages, Would that the universities had been as quick to close these sites down when the students left as they were to put them up, but it seems that many 10 year old sites are still out there, untouched, unmodified and with their content long superseded.

There has in any case been no obligation to include dates on Web pages and in cases where a date is included it may variously be the date the information was first written, the date the information was posted or the date the information was last modified. And the fact that a page bears a date does not mean that the date is genuine, nor is there any way of authenticating it.

Coverage

It is hard to determine the extent of Web coverage because of the nature of the hypertext links that join web pages and web sites together. It is not technically possible to prevent one site linking to another, and with a little skill it is possible in most cases to link to any part of any site, thus implying an authoritative and authorised link, when in fact no such tacit connection was intended by the author(s) of the site so targeted. For this reason, any one web page should in principle be evaluated independently of any other, especially when leaving the 'home' site.

Further Issues

Advertising

In printed material the distinction between advertising and information is usually made explicit, or if not immediately so the reader learns over time which parts of a publication with which he or she is familiar is likely to contain each. On the web it is not always easy to discern, especially when advertising on so many sites nowadays – as an intricate and intrinsic part of web site economics – is wholly incorporated into the page and site design.

Furthermore, while it is clear in a printed publication what is advertising and what is editorial copy, on the Web it is usually impossible to know whether the advertising and informational content are being supplied by the same person or organisation. If they are, the advertising is likely to bias the informational content, and the reader needs to be aware of this.

Customisation

What each user can see on their screen and how they see it, and what they can print or download to their computer will often vary from user to user, according to the user's platform, system and browser configuration and the software and plug-ins they have installed. Indeed the server may serve different page content according to the country of access, modem bandwidth or user-defined preferences.

Furthermore many sites are dynamic, with constantly changing content, some being randomised on user access. And even respected sites change their content regularly – for example the concept of a 'daily' newspaper doesn't exist on the Web, where 'breaking news' breaks at any time of day or night.

All these factors can mean in essence that two people can go to the same website at the more or less the same time and see very different pages, with different content, which is not helpful when you want to cite a fixed reference and lead others to it.

URL structure

Today's search engines and indexes are extremely powerful for those that have learned how to use them. Still, they are far from perfect and can retrieve Web Pages totally out of context, frequently directing the user to somewhere wholly inappropriate. To quote Huw Jarvis, "Search engine 'hits' are no indication of 'quality', nor is the numerical list any true indication of the importance of the site" (Jarvis 2003, p.209).

Sometimes you can work back through the URL, stripping off sub-directories backslash by backslash, to get to a relevant home page, but equally frequently users will be bewildered by the plethora of pop-up windows and aliased web addresses that ambush them from all sides, with the result that they end up not knowing where they are.

Site managers are frequently to blame for the inability to find previously visited and bookmarked files, or files cited in another document, and it would certainly be helpful if web managers would leave files in the same place on the site and not restructure web content every few months.

An ability to understand the way URLs are constructed and the principles of directory systems is helpful here, but even then a savvy hacker can dump a directory on some one else's system which may go undetected for a while, leaving the user believing he is on one site while in truth he is somewhere else.

Accessibility

Web Pages are notoriously unstable and not only can they become temporarily unavailable (web maintenance, too much traffic, server down, etc.) but they are likely to move or disappear without notice. Far worse, they can be altered without their owners knowing, either accidentally or intentionally. Many sites are not protected by firewalls and a competent hacker may still penetrate those that are. After all, for a skilled and dedicated hacker who can work his way into the Pentagon a University web site is hardly likely to present much of a challenge.

Online publications

Many online journals are available nowadays. Some of these are online versions of printed journals, carrying much the same content. Others, however, are wholly electronic. In this latter case, articles which are 'refereed' are far more likely to be reliable sources that those which have just arrived from nowhere to fulfil the editor's publication schedules. Check carefully to see if there is a review board and what the publishing policy of the journal is.

Caution should also be taken when quoting from electronic lists and discussion forums. Just because someone posts a message citing the words or opinions of a third party does not automatically mean that the 'quote' is accurate – it should, as in all such cases, be independently verified.

Data entry

It is also useful to reflect on how information gets into digital form in the first place. Web resources are often keyboarded or scanned. Keyboarding of long texts is often farmed out as piecework to people who may be unfamiliar with technical texts – or even in extreme cases, the language. Like humans, OCR (optical character recognition) engines have an unfortunate tendency to make sense of unfamiliar words by changing them to something within their repertoire. Spell checkers don't always help here, merely checking that each word typed corresponds to a word in the language, and not necessarily to the one intended by the author. Given this potential for error, along with the high costs of professional proofreading, it is hardly surprising that web sites have so many inaccurate texts and data.

Possible Solutions

One solution to the problem of site integrity has arrived in the shape of the portal – a web site that purports to guarantee the integrity of its content and that any outgoing links on its pages are acceptable to the site administrators. No site administrator can stop incoming links but once users arrive at the portal they know (in theory at least) where they are and what they can expect.

The down side for the user is that many such sites are passworded. There are many reasons (legal, commercial, protection of data, etc.) why academic or commercial institutions are unable or unwilling to share all their content with the world at large. Most reach a compromise and make some of their information generally available and limit other areas to their own Intranet, available only to their own staff and students, and accessible only through their own onsite terminals and with a personalised password.

Another emerging solution may lie in Extensible Markup Language (XML), which provides flexible and customisable identification of information through standardised, embedded databases. One expert in the field writes that XML "… provides a robust, non-proprietary, persistent, and verifiable file format for the storage and transmission of text and data both on and off the Web' (Flynn 2002).

XML is actually a metalanguage and, put simply, using XML means that in addition to page content a web page can also contain identification and categorisation data – in the form of data base fields - that will define its content and provenance more precisely. These data can include classification information (such as the Library of Congress subject areas), institution identification (verifying an institution as a member of one of a number of professional bodies), etc., but can also include user-defined fields for whatever purposes may serve the user(s).

It is not the purpose of this paper to discuss XML in any level of detail, and indeed beyond the competence of this author to do so,. The system is still in development, and at present is not 'foolproof', but if or when fully implemented will be a great boon to academics keen to separate the wheat from the chaff.

Other low-tech solutions include restricting a search to pre-defined domains by using a limiter (such as host: ac.uk) that will instruct the search engine to search in a limited area, in this case UK university sites.

In addition to these technological solutions, as has been said earlier, it is of course vital that users demanding data integrity apply rigorous criteria not only where they look but also in evaluating what they stumble upon on in their wanderings around the Web.

Beyond Web Texts

It has been the purpose of this paper to suggest the caution in the evaluation of printed pages on the web that one would with printed resources. However, many of the same criteria can be applied to other media that can be embedded in or called up from a web page. The content of embedded or streamed audio or video files should be given the same measure of healthy prudence, as should content of blogs, pod casts and other material of uncertain origin.

It is in fact increasingly hard to define what a 'web page' is, when so much of its content is stored in different locations. Typically, routines call on such diverse sources as scripts, includes, media files, dynamic data bases, and many others to present the user with a viewing experience that may indeed be unrepeatable, even on a screen refresh. Variables that determine what viewers actually see may include the geographical location of a user's ISP, the software configuration on the computer being used, the time of day, current server traffic, a user's previous browsing experience, and many others.

Political considerations may also be a factor controlling what a web user may and may not access. The recent entry of Google into mainland China may be a useful reference point here, although it is worth noting that as early as 2004 technology was finding solutions to provide access for Chinese web users (BBC News, 2004).

Conclusion

The Web is only one source of information, not always reliable, which can be very useful for researching certain topics on certain sites, and of little value – and possibly dangerous - for others. Judicious use of the critical faculty will improve the chances that the data accessed will be genuine and useful.

Students should be taught that to research a topic thoroughly they should use a variety of sources, both Web and non-Web. As a work-around, and if available, 'hard' material (available as CDs and DVDs), if printed by responsible institutions and reasonably up-to-date, may be more reliable than the Web in some respects. But beyond these, this paper suggests there is still very much a place for the old fashioned 'dead-tree' library, both as a source of authoritative texts and also as a means of honing study and referencing skills that might otherwise atrophy if students rely exclusively on electronic sources..

Many of the points made above are perhaps obvious to a mature, adult user of the Web. However, as the Web is increasingly being recommended for educational research projects and may - in the more developed world at least – be in the process of replacing the use of printed resources, younger users need to be warned of the limitations too. A new module in a study skills programme to this effect might not go amiss.

References

Arase, V. (2003). Message posted to "Has anyone bought Norton Systemworks 3?" In *MacUseNet* discussion Group. Retrieved 11 February 2006 from http://www.macusenet.com/archive/index-t-8972.html

BBC News [UK version]. (2006). Accessed 5 February 2006 at http://news.bbc.co.uk/1/hi/technology/3548035.stm

Beck, S. (1997). Evaluation criteria. In *The Good, The Bad & The Ugly: or, Why It's a Good Idea to Evaluate Web Sources.* Accessed 5 February 2006 at http://lib.nmsu.edu/instruction/evalcrit.html

Flynn, P. (Ed.) (2002). *The XML FAQ v. 2.1*. Accessed 15 September 2003 at http://www.ucc.ie/xml/faq.xml

Jarvis, H. (2001). Internet usage of English for academic purposes courses. In *Recall, 13*, Cambridge University Press.

Merriam-Webster's Online Dictionary. (2006). Accessed 5 Feb 2006 at http://www.m-w.com/

Biography

Martin Eayrs is a lecturer at the University of Salford, UK. He has previously worked at the University of Essex, UK, and at the Universidad de Los Andes in Venezuela. He was for seventeen years director of a private language school in Buenos Aires, and latterly language consultant to the British Council in Argentina. After twenty-three years in Latin America he returned to the UK in 2000.