Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 133-138

Does metadata count? A Webometric investigation

Alastair G. Smith
School of Information Management
Victoria University of Wellington
New Zealand
Alastair.Smith@vuw.ac.nz

Abstract

This study investigates the effectiveness of metadata on websites. Specifically, the study investigated whether the extent of metadata use by a site influences the Web Impact Factor (WIF) of the site. The WIF is a Webometric measure of the recognition that a site has on the web. WIFs were calculated for two classes of sites: electronic journals and NZ University Web Sites. The most positive correlation was found between the substantive WIF of the electronic journal sites and the extent of Dublin Core metadata use. The study also indicates a higher level of metadata use than previous studies, but this may be due to the nature of the sites investigated.
Keywords: *metadata, effectiveness, evaluation, Web Impact Factors, search engines, electronic journals, university web sites.*

Introduction

There has been much discussion of the value of metadata in providing intellectual access to digital objects. In library and information management circles the value of metadata is taken as a given. However there has been relatively little empirical evaluative investigation of the benefits of metadata. Is metadata simply a "good thing" along with motherhood and apple pie, or can its value in enhancing the value of sites, and intellectual access to them, be demonstrated objectively?

We do know that on the World Wide Web relatively few sites use metadata (Lawrence & Giles 1999). When metadata is used, it is often not used effectively. For instance a metadata template may be copied across sites without being modified to reflect the intellectual content of the site. As an example the Intellectual Property Office of NZ site (http://www.iponz.govt.nz) shares metadata with motor vehicle registry, so entry page for intellectual property office includes the inappropriate keyword "motor vehicles".

How could we evaluate the impact and benefits of metadata? Two possible approaches present themselves.

We could investigate the impact of metadata on searching: carry out an empirical investigation of the effectiveness of searches for documents which have metadata attached, and compare this the retrieval of documents without metadata. Such research needs to take account of issues relating to relevance, and evaluation of search engine performance (Oppenheim et al. 2000). In particular such research would need to choose search terms independent of the language used in the target documents, and in their metadata. Such a study has been carried out (Henshaw & Valauskas 2001), in which the retrieval of articles from an electronic journal were compared before and after the addition of metadata; the results indicated that metadata in itself did not impact on the ranking or retrieval by Internet search engines.

Another approach is to evaluate the impact factor of websites and relate this to the extent of metadata use. A Web Impact Factor (WIF) is a relatively new measure of the extent to which a site is linked to by other sites, and is analogous to a citation count in the print environment. Broadly, it is a measure of the extent of the reputation of a site, the extent to which it is linked to and recognised by other sites.

WIFs are part of the methodology of webometrics. Björneborn (Björneborn 2002) defines 'webometrics' as: "The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web, drawing on bibliometric and informetric methods".

The idea of applying bibliometric techniques to the web was proposed by Almind and Ingwersen (Almind & Ingwersen 1997). Ingwersen (Ingwersen 1998) proposed the measure of the WIF, analogous to the Journal Impact Factor in the print publishing environment. Broadly defined, a Journal Impact Factor is the ratio of citations made to a journal to the number

of citable articles in the journal. Ingwersen proposed that the WIF should be defined as the ratio of links made to a website, to the number of pages at the website. Ingwersen distinguished between:

- the simple WIF, the ratio of all links to the number of pages;
- the internal WIF, the ratio of internal links within the site to number of pages;
- the external WIF, the ratio of links made from external sites to the target site, to the number of pages at the site.

In practice the external WIF appears to be the most valid measure of impact for a site. It is noteworthy that this is similar to the Google concept of page rank (Brin & Page 1998). WIFs can be calculated from data derived from searches on web based search engines, for instance AltaVista. While most of the major search engines can in theory be used for webometric study, in practice AltaVista provides the best combination of a large database, consistent results, and boolean logic for combining complex search results. Problems with Altavista in an earlier study (Smith 1999) appear to have been overcome.

Thelwall (Thelwall 2000) has attempted to correlate WIFs with external measures of research output of British universities, and found that a WIF that concentrated on research based pages gave the highest correlation with external measures. This result was broadly confirmed in a parallel study of Australasian Universities (Smith & Thelwall 2002).

This paper describes an exploratory webometric study, attempting to establish if there is a correlation between the impact factors of electronic journals and of New Zealand University web sites; and the extent to which metadata is used on the site.

The study also tested the extent to which links made to e-journals were to the e-journal as an entity (for instance from a list of e-journals) or to specific articles or other substantive material in the e-journals (the equivalent of a print citation to a specific article).

Methodology

A number of e-journal sites were surveyed. 33 E-Journals were selected from a range of sources, including the *Electronic Journal Miner*, http://ejournal.coalliance.org/, using the following criteria:

- full text of journal articles were freely accessible on the web;
- the journals were pure e-journals, i.e. no print equivalent that could "pollute" citations;
- the journals were refereed, with at least some scholarly research articles;
- the journals had a distinctive URL that could distinguish the content of the e-journal.

For each e-journal, the following data was gathered:

- [P] number of pages spidered by AltaVista (host:{url} or url:{url} depending on whether the URL was a domain (e.g. for firstmonday.org, the host command was used) or a subdirectory (e.g. for dlib.org/dlib, the url command was used).
- [X] number of external links made to the e-journal (link:{url} and not host:{url} or link:{url} and not url:{url}).
- Proportion of pages spidered by AltaVista that contained metadata (keyword, or description) or DC metadata. This was done by sampling the first 10 URLs in the AltaVista hit list. In advanced search mode AltaVista presents results in random order, so this is a valid sample. In retrospect a more thorough study would include more URLs, but this was felt at the time to be a sufficient sample to indicate the extent of metadata use by the site. No attempt was made to judge the quality or quantity of metadata; pages were simply counted according to whether keyword or description metadata was present, and whether it was in DC format.
- [L] Proportion of linking pages that linked to substantive content in the e-journal. Many links are made to an e-journal from lists of e-journals, which does not imply impact; references made to specific articles and other content are potentially a better indication of impact.
- A similar data gathering exercise was followed for the eight NZ University websites, except that no attempt was made to assess the substantive nature of the links.

From this data, two measures of impact factor were calculated:

- The "original" external WIF, the ratio P/X.
- The substantive WIF, the ratio of links made to substantive content in the e-journal, the ratio P/X*L/100. This measure is closer to that of a journal impact factor, since it excludes links made to an e-journal from lists, which do not imply a measure of recognition.

Results

Electronic Journals

The raw data from the study is provided in appendices 1 & 2. Interestingly, comparisons using the "original" external WIF, the ratio of links from external sites to the e-journal to the number of pages at the e-journal, show little evidence that extent of metadata enhances the impact factor of the journal.

Average WIF for no metadata	6.71
Average WIF with metadata	4.27
Average WIF with DC metadata	5.33

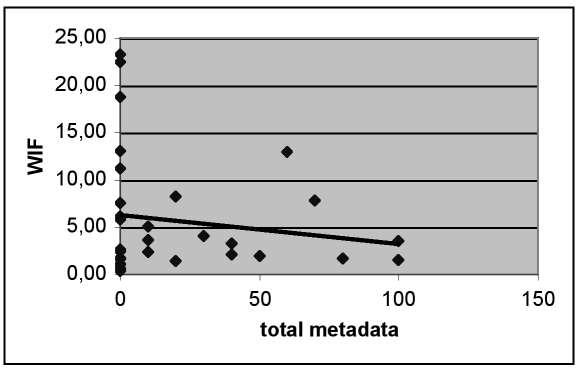

Figure 1. "Original" WIF against total metadata

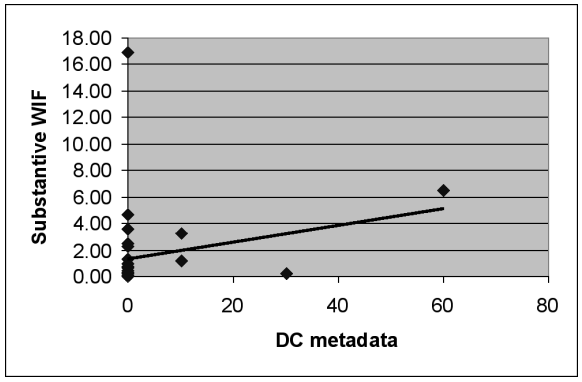

Figure 3. Substantive WIF vs DC metadata

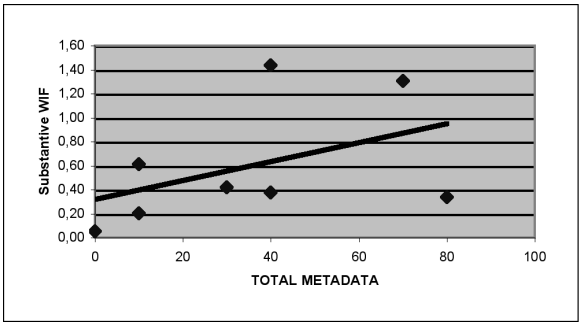

Figure 2. Substantive WIF against total metadata

A graph of the total metadata against the WIF indicates a slightly negative correlation (an Excel correlation coefficient of –0.15):

However the substantive WIF gave more support for the value of metadata. A graph of the total metadata against the substantive WIF indicates a slight positive correlation (an Excel correlation coefficient of 0.06), while a graph of the DC metadata against the substantive WIF indicates a stronger correlation (Excel correlation coefficient of 0.19):

Average subst WIF with no metadata	1.46
Average subst WIF with metadata	1.90
Average subst WIF with DC metadata	2.77

The data also allowed an estimate of the proportion of pages in e-journals that use metadata.

% pages with metadata	19.61
% pages with DC metadata	2.94

This contrasts with the Lawrence & Giles (1999) estimate of 0.3% of sites using DC metadata.

NZ University websites

A similar study was carried out of NZ University websites. No attempt was made to distinguish "substantive" links from others: this is too subjective when sites do not have clearly distinguishable information units in the way that e-journals have, and Altavista does not clearly distinguish "research pages" from other pages at the site, in the way that Thelwall's specialised webometric spider (Thelwall 2000) does. However in comparison with e-journals, a positive correlation (Excel correlation coefficient = 0.45) can be made between the external WIF and the extent of total metadata use. There is a negative correlation (Excel correlation coefficient = -0.21) with the extent of DC metadata use, but this may be because of the small amount of DC metadata in the sample. The proportion of pages with metadata were similar to those for the e-journals.

% pages with metadata	16.86
% pages with DC metadata	4.35

Discussion

Perhaps most significant was the relatively small amount of metadata use found in the study. Even in Library and Information Management e-journals, metadata was relatively rare; to the extent that at least one article on the topic of metadata did not include metadata in the HTML header (Caplan 1995).

Is use of metadata increasing? The increase between the Lawrence and Giles figures metadata (Lawrence & Giles 1999) and those found in this study are encouraging; however use of metadata by university sites and electronic journals would be expected to be higher than the norm. On the other hand, perhaps we don't want metadata to be too widely used: to some extent metadata acts as a filter, so that material that is worth retrieving will have metadata added, while more transitory material will not have metadata attached.

The study does demonstrate that the amount of metadata attached to a site influences at least some measures of the impact of a site. The correlation between the amount of Dublin Core metadata in elec-

tronic journal sites and the substantive external WIF is the most positive. For electronic journals, there is a slight negative correlation between the amount of metadata use and the standard external WIF; this may indicate the lack of validity of the standard external WIF as an impact measure for electronic journals, since this measure does not distinguish between links to the electronic journal as an entity, and links to substantive content. For NZ University sites, there is a positive correlation between the total metadata use and the impact factor of the site.

While these results are mixed, they are encouraging, given the effort expended on defining metadata standards. We may be approaching a critical mass of metadata, where metadata is sufficiently widely used in certain contexts to achieve usefulness, and will be adopted by search engines. According to Sullivan (Sullivan 2002), meta description tags are utilised by all major search engines except Google; meta keyword tags are utilised by Altavista and Inktomi, but not by FAST and Google.

This preliminary research does not positively confirm the value or otherwise of metadata. It indicates the need for further research to confirm the results of this exploratory study. In particular larger samples could be used to confirm the extent of metadata use by target sites. Larger numbers, particularly of university/research sites, and other classes of sites could be studied. The effect of quality and quantity of metadata used could also be studied.

References

[URLs checked 13 June 2002].

Almind, T.C. & Ingwersen, P. 1997, 'Informetric analyses on the World Wide Web: methodological approaches to "Webometrics"', *Journal of Documentation,* vol. 53, no. 4, p. 404–426.

Björneborn, L. 2002, 'Defining webometrics [Message to webometrics@coombs.anu.edu.au 30 May 2002]'.

Brin, S. & Page, L. 1998, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Available: [http://www-db.stanford.edu/~backrub/google.html].

Caplan, P. 1995, 'You Call It Corn, We Call It Syntax-Independent Metadata for Document-Like Objects', *Public-Access Computer Systems Review,* vol. 6, no. 4. Available: http://info.lib.uh.edu/pr/v6/n4/capl6n4.html

Henshaw, R. & Valauskas, E. J. 2001, 'Metadata as a catalyst: experiments with metadata and search engines in the Internet journal, First Monday.', *Libri,* vol. 51, no. 2, pp. 86-101.

Ingwersen, P. 1998, 'Web Impact Factors', *Journal of Documentation,* vol. 54, no. 2, p. 236–243.

Lawrence, S. & Giles, C.L. 1999, 'Accessibility of information on the Web', *Nature,* vol. 400, pp. 107-9. Available: summary at http://www.wwwmetrics.com/.

Oppenheim, C., Morris, A., McKnight, C. & Lowley, S. 2000, 'The evaluation of WWW search engines', *Journal of Documentation,* vol. 56, no. 2, pp. 190-211.

Smith, A.G. 1999, 'A tale of two web spaces: comparing sites using web impact factors', *Journal of Documentation,* vol. 55, no. 5, pp. 577-92.

Smith, A.G. & Thelwall, M. 2002, 'Web Impact Factors for Australasian universities', *Scientometrics,* vol. 54, no. 3, p. 363–380.

Sullivan, D. 2002, *Search Engine Features For Webmasters*, Available: [http://www.searchenginewatch.com/webmasters/features.html].

Thelwall, M. 2000, 'Extracting Macroscopic information from web links', *Journal of the American Society for Information Science & Technology,* vol. 52, no. 13, pp. 1157-1168.

Appendix 1. Electronic Journals

Name	URL	Pages [P]	External links [X]	% metadata (keyword/ description)	% DC metadata	% total metadata	% links to articles [L]	WIF	Substantive WIF
LibRes: Library And Information Science Research Electronic Journal	libres.curtin.edu.au	38	67	0	0	0	0	1.76	0.00
Electronic Journal of Radical Organisation Theory	mngt.waikato.ac.nz/ejrot	46	122	0	0	0	0	2.65	0.00
Canadian Journal of Educational Administration and Policy (CJEAP)	umanitoba.ca/publications/cjeap/	22	128	0	0	0	0	5.82	0.00
Electronic Journal of Probability	math.washington.edu/~ejpecp	130	982	0	0	0	0	7.55	0.00
Electronic Transactions on Numerical Analysis (ETNA)	etna.mcs.kent.edu	87	977	0	0	0	0	11.23	0.00
Earth Interactions: An Electronic Journal Serving the Earth System Science Community	earthinteractions.org	12	279	0	0	0	0	23.25	0.00
EM Electronic Markets	electronicmarkets.org	605	934	100	0	100	0	1.54	0.00
Cybermetrics	cindoc.csic.es/cybermetrics	81	288	100	0	100	0	3.56	0.00
Architronic	architronic.saed.kent.edu	292	110	0	0	0	20	0.38	0.08
Journal of Information Law and Technology	elj.warwick.ac.uk/jilt	2,325	1,298	0	0	0	30	0.56	0.17
Electronic Journal of Biotechnology	ejb.org	245	417	80	0	80	10	1.70	0.17
E Law - Murdoch University Electronic Journal of Law	murdoch.edu.au/elaw	1545	972	0	0	0	30	0.63	0.19
Information Research	informationr.net/ir	119	230	20	30	50	10	1.93	0.19
Crossings Electronic Journal of Art and Technology	crossings.tcd.ie	22	23	0	0	0	30	1.05	0.31
Electronic musicological review	cce.ufpr.br/~rem	189	307	0	0	0	20	1.62	0.32
Folklore: An Electronic Journal of Folklore	haldjas.folklore.ee/folklore	462	518	0	0	0	40	1.12	0.45
International Electronic Journal for Leadership in Learning	ucalgary.ca/~iejll	71	435	0	0	0	10	6.13	0.61
Reading Online	readingonline.org	805	1697	40	0	40	30	2.11	0.63
E-Journal	hanover.edu/philos/ejournal	140	150	0	0	0	60	1.07	0.64
Interactive multimedia electronic journal of computer-enhanced learning	imej.wfu.edu	247	356	20	0	20	50	1.44	0.72
Australasian Journal of Disaster and Trauma Studies	massey.ac.nz/~trauma	76	279	10	0	10	20	3.67	0.73
Screening the Past: An International Electronic Journal of Visual Media and History	latrobe.edu.au/www/screeningthepast	349	855	0	0	0	40	2.45	0.98

Appendix 1. Electronic Journals (continued)

Name	URL	Pages [P]	External links [X]	% metadata (keyword/ description)	% DC metadata	% total metadata	% links to articles [L]	WIF	Substantive WIF
Journal of World-Systems Research	csf.colorado.edu/jwsr	51	259	10	0	10	20	5.08	1.02
Ariadne	ariadne.ac.uk	1073	2556	0	10	10	50	2.38	1.19
Journal of Digital Information	jodi.ecs.soton.ac.uk	270	880	40	0	40	40	3.26	1.30
The Journal of Library Services for Distance Education	westga.edu/~library/jlsde	40	523	0	0	0	10	13.08	1.31
Interpersonal Computing and Technology; An Electronic Journal for the 21st Century	jan.ucc.nau.edu/~ipct-j	18	404	0	0	0	10	22.44	2.24
Journal of Electronic Publishing	press.umich.edu/jep	244	2,007	20	0	20	30	8.23	2.47
Dlib Magazine	dlib.org/dlib	882	3581	20	10	30	80	4.06	3.25
Essays in History	etext.lib.virginia.edu/journals/EH	69	411	0	0	0	60	5.96	3.57
Journal of Computer-Mediated Communication	ascusc.org/jcmc	322	2524	70	0	70	60	7.84	4.70
First Monday	firstmonday.org	157	2030	0	60	60	50	12.93	6.46
Public-Access Computer Systems Review (PACS Review)	info.lib.uh.edu/pr	49	920	0	0	0	90	18.78	16.90

Appendix 2. NZ University websites

Name	URL	pages	external links	% metadata (keyword/ description)	% DC metadata	% total metadata	WIF
Massey	massey.ac.nz	213,151	10,824	0	0	0	0.05
Auckland Univ of Technology	aut.ac.nz	10,355	2,112	10	0	10	0.20
Otago	otago.ac.nz	32,662	11,039	80	0	80	0.34
Waikato	waikato.ac.nz	32,639	12,251	10	30	40	0.38
Canterbury	canterbury.ac.nz	32,408	13,588	20	10	30	0.42
Auckland	auckland.ac.nz	35,431	21691	0	10	10	0.61
Lincoln	lincoln.ac.nz	3,011	3,936	70	0	70	1.31
Victoria	vuw.ac.nz	21,781	31,303	40	0	40	1.44