

Make the Web work for you

=====

by Gordon Woolf*

Everyone, it seems, will design you the best web site in the world, and, if you seek to do it yourself, there are dozens of books in the computer section of most bookshops.

But what if, instead of a flashy, Java or Active-X web site with sound, vision and animation, you just want a web site which works and brings in customers or prospective customers?

Suddenly the choice is limited.

There's an IBM ad on TV which takes this stance -- the web designer who asks the boss whether he wants a rotating graphic or one which bursts into flames. Asked for one which links an order to a stock database, he doesn't know how to do that.

The implication is that you can pay IBM for the answer.

Well that's one route, but the information is out there, if only you can find it.

Firstly, the web space market has been changing.

Most Internet Service Providers continue to offer "free web space", but, if you enquire further, you will find it is with severe limitations. You see, it is not how much you use your account -- if you have a business to run, you are unlikely to reach the free allocation your ISP will give you.

The real cost to them comes when you have a site which is being contacted in large numbers by web users around the world. And especially if you have many files on offer. The PDFs (Acrobat files) we offer may only be 45kb a time, but what if a thousand people download each of them every month? Or every week?

If your ISP is linked to Telstra, say, that's an awful lot of megabytes they have to pay Telstra for moving around.

So accept that for running a business web site, you are moving beyond the free site. You may find some amazingly good offers. but wait for the complaints that people can't link to your site, or get broken downloads.

Your own dialup contact with your ISP may seem first class, but you may be in for a shock if you ask for a report from one of the web sites which provide reports on how easy it is to contact your site. You can also ask business contacts, relatives and friends around the world for their comments.

Take note of sites you visit which seem fast and friendly. Unless they are a major corporation it is highly likely they are hosted by an ISP, and the computer world is still friendly enough that if you ask the webmaster, you'll probably get an answer.

If you have a site which is likely to attract significant custom (I refuse to refer to "hits", and I'll say why in a minute), they will lose interest if you are not going to have your pages designed by their resident design service (and note that they nearly all call them "design" services).

One reason is that many small ISPs are a computer or two connected directly to the site of a larger provider -- and they have to pay that provider by the megabyte even though they are charging you by the hour. If too many people log on to your site, they can afford a loss on that, provided you are paying them for maintaining the site.

This is not unreasonable. But it means the do-it-yourself web creator has to beware.

Other limitations to watch for include the ability to load what are called CGI scripts. These are "Common gateway interface" scripts, a bit like batch files (like the old DOS days) -- they tell the server what to do, and are, like batch files, written in something which resembles English. It still pays to use well tested "library" files to control the real processing, but you will want to be able to change the script which calls the library -- even for something as simple as changing the subject lines and text of emails sent from the web site or to change the text of a web page created by the script to send back to the browser to confirm that their enquiry has been sent to you.

Many service providers set conditions on CGIs -- yes you can use them, but only the ones they provide, or they may charge a fee for giving yours the once over before they allow them to be uploaded (or they'll upload them on your behalf to a directory where you do not have access).

Others take the view that what they are providing is processor time -- they'll include an allocation to you per day, or per week, and what you do with it is up to you. If it goes over because of a badly written CGI script, well that's your problem.

The time limit they quote may seem impossibly short -- a few seconds a day -- but remember this is the CPU time on a fast processor, and you can do a lot of processing in a single second. Our ISP says we can have 1000 cpu-seconds a day (a little over 15 minutes), of which I'm currently using 3 or 4. You can do a lot in a cpu-second. Their estimate is that in 1000 seconds I should be able to run up to 100,000 cgi scripts (for forms retrieval, feedback details, page counters etc.). If we ever get to that, I'll retire on the proceeds of the orders.

Those sites which give you greater access are also usually the ones which do the least handholding. They'll give you some FAQs (or guide you to FAQs at other sites), but it is up to you to telnet into your own directories and set the access levels using raw UNIX commands. It's even more like the old DOS days, though some of the top-of-the-range website management software can help if you can afford it.

You also have to consider what your next step will be if your web page is successful and you start to get a flow of enquiries. We make it clear to anyone who fills in our order form that we are not on a secure server, and that while we are happy to take credit card details, we suggest they may like to fax their order, or send the details split between separate form and email submissions, or even encode it with PGP (Pretty Good Privacy).

However, we did check that as orders build we will be able to migrate to a secure server without any apparent change in our web address. It will cost us more, but I hope the day when that is warranted will not be too far off. At least we have planned for it. Changing servers without your own domain address can be a pain. You may discover, as we did, that your old web page is listed on the CDs of web pages which are popular with schools and educational centres as a means of letting people browse offline.

Another limitation of many ISPs is the kind of report they will provide. I used to think it was good when I got a list of "hits" (and I said I'd return to that) but then I found out what a hit was. Every little graphic which loads is a hit, so one person glancing at one page and passing on can count as maybe 5 to 10 "hits".

Then I started to get reports on how many such hits there had been on each file, so I could start to work out how many people had really been looking. More interestingly, there came a report of what countries those hits had been from, and I started to see that down at the bottom of the list, in contacts from Bulgaria or Lithuania (and some countries I had to go to the atlas to find), they were mostly around the 6 hits. As it seems most unlikely that more than one person in a week was calling from Macedonia or Paraguay, I could see a pattern that one look at one

page equalled from 5 to 7 hits.

Now I get full access logs which can be looked at raw or decoded by one of the various web statistics programs, either on your provider's computer or your own. I started with a basic report as suggested by the provider, but now, every Sunday I download the week's logs and run a copy of Analog <www.statslab.cam.ac.uk>, the least graphic but most informative of the statistics programs.

This offers a huge range of options, such that I now have a command line to run Analog which reads as follows:

```
analog -h to +frr1 +t +Sr-250 +c +e -p -u- -q -G +C"GRAPHICAL OFF"  
+nworsleypress.com httpd_access.*
```

I won't go into a long explanation, and it is probably in a not very logical order because it has grown as I've found out what is available and how to specify it. Suffice it to say that it is a batch file run from an icon in a folder on my desktop.

The report is an HTML file (what else?). There are extracts here of the kind of information, but one of the most interesting is what you get when you specify ALL the referring sites. These are the sites which a person was at when they clicked on a link to come to your site.

Of course many of them are just other pages on your own site, and it will show which friendly sites have links to yours that people are actually using. However, to me the most useful information came in the details you will never get unless you specify it yourself from the logs: the sources of the one-time connection. A large proportion of these turn out to be from search engines, and, because many search engines put a complex internal command into their address line, you can see what people typed in to make their search before they clicked on one of the results to come to you.

For example, a lot of people came to our site by typing in queries like "how to publish a book" or "newspapers + publishing" or "pagemaker hints". That shows our efforts to get suitable listings on search engines are beginning to work. However, I can't explain how someone got referred to our site by typing in "hypogonadism+irritability+symptoms". Can Excite.com please explain?

Getting entries in the search engines is worth an article in itself, but, once you get a web site you will get plenty of emails offering to submit your site to 50 or even 250 search engines free of charge, or for a "very reasonable" \$29.95, US dollars of course.

Don't do it. It may take you ages to go to Yahoo and Excite and WebWombat (the very good Australasian-only search engine) and all the other general and specialist engines, but you can tailor your entry to each site after having a look around to see how they work. And you can stay connected to check that they accepted your submission. It will take time.

Some will not ask for any information: just a web address. These are the ones which will send a robot to look at your site, and it is important to have it ready for their visit.

Remember that a site which is not linked to from another site, will never be found -- by anyone. The web works by redirection, and the robots (often called spiders -- what else would roam a web?) follow links, taking, in many cases, only the first few words they find as their guide.

The most important words are the title: those words which appear in the title bar of the browser. It should be the most considered short phrase you have ever written because it is the phrase that matters most in how you come up in most search engines.

The next are an often overlooked series of what are called Meta tags -- tags which mean nothing in getting your HTML code onto a browser screen but which can mean a lot in how other people reach that page. Firstly <META NAME="description" CONTENT="the words which come here will appear in many search engines as the description of your page">
Then come <META NAME="keywords" CONTENT="sequence, words, people, likely, search, under">

And it doesn't do any harm to repeat these under a comment tag <!-- -->. But don't over do it. The robots are getting quite clever at realising they're being conned if they find the same word turning up a dozen times in quick succession.

However, the number of times a searched word turns up in a page will often control the order in which your page appears in the sequence presented as the result of a search.

Remember too that the ISP you use to provide your web space need not be the one you dial in to for email and general access. It need not even be within local call range, as you can dial in to them via your local ISP, and transfer all files by FTP or telnet in via the Internet itself. Your email and other personal services can, for instance, still be via your trusty MelbPC connection -- you can log in locally to retrieve email from a server across Melbourne or across the world.

And where did we find this information that has helped us to a position of getting a small but steady flow of orders from the web? On the web itself, of course. It's all there. You just have to look, and keep looking. Use the search engines because in doing that you'll be teaching yourself how they work.

We have a long way to go yet. And yes, design is also important, and our pages could certainly do with some more work in that direction, but there is a lot more than design to getting yourself a working web site.

=====

*Gordon Woolf runs a small book publishing and newspaper page production and training business that, from its web site, appears bigger. You can contact him at <gordon@worsleypress.com> or via <www.worsleypress.com>

=====

```
[17/Jul/1998:06:49:10 -0700] "GET /np-pm65.gif HTTP/1.0" 200 3694
"http://www.worsleypress.com/newsletter.htm" "Mozilla/4.05 [en] (Win95; I ;
Nav)"sub13-133.stetson.edu /webadmin/home/pub.w/worsley/public_html -
[17/Jul/1998:06:49:10 -0700] "GET /vine.gif HTTP/1.0" 200 10694
"http://www.worsleypress.com/newsletter.htm" "Mozilla/4.05 [en] (Win95; I ;
Nav)"sub13-133.stetson.edu webadmin/home/pub.w/worsley/public_html -
[17/Jul/1998:06:49:10 -0700] "GET /newsltr7.gif HTTP/1.0" 200 18249
"http://www.worsleypress.com/newsletter.htm" "Mozilla/4.05 [en] (Win95; I ;Nav)"
sub13-133.stetson.edu /webadmin/home/pub.w/worsley/public_html -
```

Fig.1: An example of the raw data you leave behind if you visit a web site. It shows what you clicked on, where you came from, even what browser you were using and what server you came from. Contrary to popular myth, it doesn't show your email address.

=====

```
Total successful requests: 6 977
Average successful requests per day: 225
Total successful requests for pages: 2 013
Average successful requests for pages per day: 65
Total failed requests: 168
```

```
Total redirected requests: 3
Number of distinct files requested: 67
Number of distinct hosts served: 1 023
Number of new hosts served in last 7 days: 227
Corrupt logfile lines: 48
Total data transferred: 225 816 kbytes
Average data transferred per day: 7 291 kbytes
```

Fig.2: An example of the minimal report you should expect. Anything less is useless.

```
=====
#reqs: %bytes: domain
-----
1649: 24.25%: .net (Network)
1682: 22.87%: .com (Commercial, mainly USA)
1265: 16.57%: [unresolved numerical addresses]
295: 4.79%: .edu (USA Educational)
393: 4.44%: .au (Australia)
161: 2.67%: .de (Germany)
181: 2.57%: .ca (Canada)
61: 2.16%: .nl (Netherlands)
77: 1.80%: .fi (Finland)
44: 1.73%: .dk (Denmark)
168: 1.66%: .uk (United Kingdom)
52: 1.19%: .it (Italy)
63: 0.97%: .org (Non-Profit Making Organisations)
24: 0.96%: .br (Brazil)
80: 0.91%: .es (Spain)
29: 0.77%: .th (Thailand)
51: 0.76%: .pt (Portugal)
23: 0.76%: .ke (Kenya)
45: 0.76%: .sg (Singapore)
35: 0.55%: .at (Austria)
21: 0.54%: .yu (Yugoslavia)
45: 0.51%: .jp (Japan)
23: 0.44%: .ch (Switzerland)
11: 0.44%: .il (Israel)
12: 0.44%: .ua (Ukraine)
```

Fig.3: A typical domain report

```
=====
#reqs: %bytes: file
-----
705: 2.96%: /vine.gif
692: 1.00%: /np-pm65.gif
493: 1.20%: /newsletter.htm
435: 3.14%: /newsltr7.gif
336: 5.15%: /newsltr7.pdf
275: 0.60%: /
258: 2.50%: /pwc.jpg
254: 0.25%: /divider.gif
249: 24.66%: /np-web.pdf
249: 0.13%: /dot.gif
248: 0.46%: /format1.gif
244: 9.83%: /newsltr6.pdf
234: 22.26%: /format1.pdf
```

Fig. 4: The start of a request report. Top of the list is the background gif we use, then a gif which appears on several pages. Third comes the most viewed page --

which, incidentally, isn't our main or index page.

=====

```
1: http://search.excite.com/search.gw?search=pagemaker+plugins
1:
http://www.hotbot.com/default.asp?MT=pagemaker+tips&submit=SEARCH&SM=MC&DV=7&RG=.com&DC=100&DE=2&_v=2&OPs=MDRTP
1: http://ww2.altavista.digital.com/cgi-bin/news?msg@17722@alt.aldu.s.pagemaker
1:
http://netfind.aol.com/search.gw?search=how+to+small+press+publish&lk=excite_netfind_us&nrm=n&pri=on&xls=b
1: http://www.hotbot.com/?_v=2&OPs=MDRTP&MT=Pagemaker&act.next.x=2&act.next.y=9
1: http://ink.yahoo.com/bin/query?p=adobe+pagemaker+templates&hc=0&hs=0
1: http://search.excite.com/search.gw?s=pagemaker+AND+booklet+AND+additions&trace=L
1:
http://search.excite.com/search.gw?c=web&s=hypogonadism+irritability+symptoms&showSummary=true&start=40&perPage=10&next=Next+Results
1:
http://www.hotbot.com/?MT=PageMaker+Bulgaria&submit=SEARCH&SM=MC&_v=2&OPs=MDRTP&base=40
```

Fig. 5: A small part of the referrer report: what people asked the search engines.