



# Information Overload

Issue 20, April 2004

---

## Welcome:

Welcome to this month's issue of Information Overload. I would like to thank those of you who took the time to comment on the White Paper on electronic archiving. As promised we will be publishing those comments in a later edition, so if you would like to add your voice to the melting pot that is electronic archiving, we would love to hear from you.

So how on earth do we follow an epic such as the white paper on electronic archiving? Well to be honest we can't, so we have changed direction as is our want, and this month we are looking at the issues surrounding the use of the Internet as a search tool.

As always, if you would like to see us cover any other topics, we would love to hear from you. Just send an e-mail to [training@iea.com.au](mailto:training@iea.com.au). We would like to thank you in advance for forwarding this edition to friends and colleagues.

We hope you enjoy reading, have a great week.

---

## In this Issue we will be looking at:

- Searching the World Wide Web
- Milestones in Search Engine Development
- The Invisible or Deep Web
- Search Engine Optimisation
- Bad Usability Equals No Customers
- A Thought to Ponder

## Searching the World Wide Web

It is said that even the most prolific search engines only index a very small percentage of the pages that make up what we know as the Internet, World Wide Web or what has been termed – the “surface web”. Google is said to index some 4 billion pages and appears to be the search engine of choice by most searchers. Unfortunately there are another 6 billion pages that Google does not currently index. (2004 Bright Planet Corporation), which is probably a good reason not to rely on Google for all your searching needs.

So how do you search multiple search engines without having to remember individual search engines? You will be pleased to know that there are such things as metasearchers. These are the search engines of search engines and come with interesting names such as Dogpile and Metacrawler. These are able to search multiple search engines at the same time, and

are able to deliver a much wider range of results to any given query.

However, as we will see, the real problem faced by most search engines is that they cannot index what they cannot see. To be found by a search engine, a web site, web page or database either has to be submitted by the author for inclusion in a search engine listing (usually for a fee), or they are harvested with the use of robots or spiders. These crawl through the pages, following links, much like a human might search the Internet. However, in order to be found, a web page has to be written in static HTML code or have enough static HTML code to be visible to the spiders, plus have enough links and keyword rich content to merit inclusion in the search engine directory. These pages are then ranked in order of a search engines algorithm.

As you can imagine, this raises a few questions. Why aren't some sites indexed? Why do some queries give lots of useless results? Is there a better way to search the Internet?

As we mentioned previously, sites have to be written in HTML code in order to be found, they then need enough keyword rich content for a search engine to decide whether or not the site is worth indexing. This is where Search Engine Optimisation comes into play,

---

© IEA 2004. All rights reserved. You are free to use material from the Information Overload newsletter in whole or in part, as long as you include acknowledgement of source.

and we will be looking at the various ways an organisation can leverage this in order to gain better ranking within a particular search engine directory.

However, with the advent of JavaScript and dynamically generated content, most web sites do not generate pages in static HTML code and are not found by the robots or spiders during their daily crawl through the web.

The question is why are web sites not being written in normal HTML so they can be found?

Dynamically generated content such as directory listings, exist only in a database, and the "results" only construct when you have asked a direct question of the sites own search engine. If you have ever used online telephone directories ([whitepages.com.au](http://whitepages.com.au)) you will know what I mean. Databases such as these are known as "the invisible web" or "Deep Web" and are rarely if ever found within the surface web domain of the search engines.

As with all things, this can change given time – but the basic algorithms that make up the traditional search engines will have to change dramatically in order to harvest this type of content. Alternatively, web developers will have to give enough content in static HTML so they can be found within the surface web domain of the search engines.

Search engines have made a rapid rise in their use and popularity since the early days of development, as the following list shows.

## Milestones in Search Engine

### Development

**1990 – Archie:** Alan Emtage, a student at McGill University in Montreal, Canada, invents the internet's first search engine. FTP owners post information about files stored on their FTP servers scattered across the internet and indexes them so users can access them.

**Feb 1993 – Architext:** Six Stanford U/G students start what will become one of the web's most popular search engines – Excite

**June 1993 – Wanderer:** Matthew Gray starts a controversy over indexing methods when he creates the WWW wanderer, an automated search and indexing system. The Wanderer's early forays cause a netwide

slowdown as it attempts to access site hundreds of times each day.

**Jan 1994 – Infoseek:** Founded by Steven Kirsch

**Jan 1994 – Yahoo:** Two Stanford University students – Jerry Yang and David Filo create **Yet Another Hierarchical Official Oracle**

**April 1994 – Webcrawler:** Brian Pinkerton, a computer science student at the University of Washington, begins a small project that grew out of a seminar discussing the growth of the WWW.

**May 1994 – Lycos:** Michael Mauldin, building on work by John Leavitt, starts Lycos (Latin "Lycosidae", "Spider")

**1995 – Google:** Sergey Brin and Larry Page meet at Stanford University. By the end of the year, they have developed "Backrub" the foundation for the Google search engine.

**Feb 1995 – Infoseek:** Infoseek search launched

**Feb 1995 – Metacrawler:** Eric Selburg, a masters student at the University of Washington creates the net's first crawler to collect all searches on the same page.

**July 1995 – Alta Vista:** Harvesting robot, "Scooter" first crawls the net.

**Sept 1995 – Inktomi:** Founded

**Oct 1995 – Excite:** Launched

**Dec 1995 – Alta Vista:** Launched

**1996 – Ask Jeeves:** Founded

**May 1996 – Hotbot-Inktomi:** Fastest crawling spider – able to index 10 million pages a day.

**Oct 1996 – Looksmart:** Australia's biggest directory of website starts.

**1997 – Go to:** First search engine that auctions page rank positions. Later changed name to Overture and bought by Google.

**Aug 1997 – Northern Light:** Starts

**Apr 1998 – Google:** Introduced by Brin and Page in a research paper entitled The Anatomy of a Large-Scale Hypertextual Web Search Engine

**Sept 1998 – MSN Search:** – Introduced

**Sept 1999 – Google:** Launched

**Apr 2000 – Teoma:** Created

*Taken from The Power of Search by Nathan Cochrane, April 13, 2004.*

<http://www.smh.com.au/articles/2004/04/12/1081621875197.html>

## The Invisible or Deep Web

If it's invisible how can I find it? The term Invisible web is actually inaccurate as the information is not invisible per se, it's just not

visible to the traditional search engines. As we have discussed the traditional search engines are not able to find information contained in dynamically generated web sites.

If you liken the Internet or World Wide Web to an iceberg, then it will come as no surprise that what you see when you trawl through the hundreds of "hits" you get when you use a search engine, is nothing compared to what is available – if only you knew where to find it. It is said that the deep web is 500 times the size of what we know as the "Internet" or visible web.

(<http://brightplanet.com/technology/deepweb.asp>)

Whilst a lot of the deep web sites are already familiar to you, sites such as Amazon, msn, yahoo, CNN, telephone directories and so on. There are a lot of other sites that you may not have heard about. Before you throw up your hands in despair, you will also be pleased to know that there are a number of gateways to the invisible web that you can try. Whilst this is not a complete listing, as with all things web based, the number of directories, listings and databases is in a constant state of flux. We hope that the sample of invisible web resources will give you a taste of what is really out there.

#### **Librarian's Index to the Internet**

<http://lii.org>

One of the most popular entry ways to the invisible web. This annotated and searchable index of over 11,000 resources has been selected and evaluated by librarians.

#### **Resource Discovery Network**

<http://www.rdn.ac.uk>

Subject "hubs" cover topics such as Life Sciences, Law, Engineering, Health and Medicine.

#### **Profusion**

<http://www.profusion.com> (formerly known as <http://www.invisibleweb.com>)

A search engine style collection of over 10,000 databases.

#### **Complete Planet**

<http://aip.completeplanet.com>

Developed by Deep Web experts Bright Planet, gives access to over 70,000 searchable databases and speciality search engines.

#### **Invisible Web Net**

<http://www.invisible-web.net>

This is a hand compiled listing of entry points to the Invisible web and was created by the

sites authors as examples for their book "The Invisible Web"

## **Search Engine Optimisation**

If your website has been created using the latest JavaScript technology and does not exist in a static environment, then chances are you will not be found using the traditional search engines. As we have already discussed – the majority of information resides in databases that can only be accessed via specific queries. So what do you have to do if you want your web site to be picked up by the spiders as they crawl through cyberspace? Simple - you have to give them something to chew on when they get to your web address. If a spider cannot see anything worthy on your web site, it will quickly head off to find other sites that it considers are worthy of their attention.

So how do you optimise your site so that it can be found? And more importantly hang around long enough for you to "sell" whatever it is you are promoting.

**Meta Description Tags** – The meta description tag is used to assist those search engines which are meta capable in summarizing your web site (note – Google does not search meta tags). The size of the meta description tag should be under 200 characters. It should describe your web site, and follow on from your <title>. For example:  
<title>The Drama Queen - Coral Drouyn, Australia's TV Drama and Soap Monarch</title>  
<meta name="description" content="Screen writer for popular Australian TV Drama and Soaps.">

In addition to the tags, you should also build a **summary paragraph** into the web page, which can be found by those search engines that do not use meta tags. This summary should be 200 characters or less and be the first visible text on the screen when the web page loads. Engines like Northern Light and Lycos use this information for their summary of your web page.

**Meta Description Keywords** – These follow the meta description tags and should be relevant to each, individual page. This allows you to draw people into different parts of your web site. If all your tags are the same, you are severely limiting your chances of being

found and indexed. Add some commonly misspelled variations. Not everyone can type or spell correctly.

If you are not sure if your web page has this information, you should right click on an area of white space on your web site and "view source" - the meta tags and meta keywords should be the first items you see.

Carrying on with the The Drama Queen example, the meta keywords are:

```
<meta name="keywords" content="drama queen, the drama queen, drama queens, coral, coral drouyn, coral drawn, coral druin, writer, scriptwriter, script writer, screen writer, screenwriter, home and away, home & away, summer bay, pacific drive, prisoner, tv shows, tv drama, tv soap, t.v. shows, t.v. drama, t.v. soap, television shows, television drama, television soap, writing for tv, blue heelers, neighbours, australian tv, australian t.v., australian television, writers block, writing academy, screen dreams, screendreams, script writing for tv drama, elk, elk consultants, elk consulting, elk consultant, e.l.k">
```

Please bear in mind that not all your pages may have content that needs indexing – especially if the navigation is poor and you give the users no option but to leave the site.

Make sure **your links are not broken**. You want the spider to be able to crawl easily through your site.

**Title tags** – All search engines use title tags to gather information about your web site. Therefore it makes sense to have relevant title tags on each page. Another case of do not be lazy when building your site. A word of caution, make sure that the words, keywords and phrases that you use are found within the body of your web site. Otherwise the robots and spiders may not rank your site as highly as it should.

**Keyword repetition** – this is the domain of the spammers. Do not use the same word twice in the same row in the tag, even if you are using different variations eg caps, plurals, different tenses and so on.

**Keyword density** – is the ratio of keywords to the total number of words on a web page. In other words, it is the number of a particular word appearing in all the different locations

(such as tags – meta, comment, alt and header tags – and the body text) divided by the total number of words. (Ideally 3-8%)

**Index Spamming or Spamdexing** – But don't try and fool the search engines – they've picked up on this one already. The over repetition of a word or words can cause some search engines to reject your site, ban your IP or penalise it by giving it a much lower ranking.

**Text the same colour as the background** – again you spammers out there need to be aware that the search engines have picked up on this one and will now relegate your web site to the spam category.

**Link Popularity** – the more sites that link to yours increases the chance of higher ranking in the search engines.

What has all this got to do with how we search the internet? Well unless you or your web developer have taken the time to ensure that your titles/tags/keywords and the keyword rich content that your site contains are correct and optimised for other people to find, then you have probably wasted a lot of time and a lot of money on something that cannot be found. If you find that you are not getting the kind of traffic that you expected to get, then chances are at least one of your components is missing or poorly utilised.

Oh and the web site that we have used in this example comes from <http://www.thedramaqueen.net>

### **Bad Usability Equals No Customers**

"In the network economy, the website becomes a company's primary interface to the customer. Indeed, for e-commerce companies the site IS the company. The user interface becomes the marketing materials, storefront, store interior, sales staff, and post sales support all rolled into one. In many cases the site even becomes the product itself. Thus having bad usability is like having a store that is on the 17<sup>th</sup> floor of a building (so nobody can find it), is only open Wednesday's between 3 and 4 o'clock (so nobody can get in), and has nothing but grumpy salespeople who won't talk to the customers (so people don't buy too much)"

*Nielsen, Jakob: Designing Web Usability, p 14*

---

© IEA 2004. All rights reserved. You are free to use material from the Information Overload newsletter in whole or in part, as long as you include acknowledgement of source.

---

## A Thought to Ponder:

If the automobile had followed the same development cycle as the computer, a Rolls-Royce would today cost \$100, get a million miles per gallon, and explode once a year, killing everyone inside.

**Robert X. Cringely, InfoWorld magazine**

---

Your comments and suggestions on the subject of this newsletter are most welcome. Or if you would like to see other issues covered in future editions, please email [training@iea.com.au](mailto:training@iea.com.au)

Please feel free to pass on this newsletter to your colleagues' friends and associates. To subscribe they should send an e-mail to [training@iea.com.au](mailto:training@iea.com.au) with "subscribe newsletter" in the subject line.

If you would prefer not to receive this newsletter, please send an email to [training@iea.com.au](mailto:training@iea.com.au) with "unsubscribe newsletter" in the subject line. If you have any suggestions as to what should be included in future editions, then please send an email to [training@iea.com.au](mailto:training@iea.com.au).

---

© IEA 2004. All rights reserved. You are free to use material from the Information Overload newsletter in whole or in part, as long as you include acknowledgement of source.

**Information Enterprises Australia Pty Ltd**  
Unit 4, Upper Level, 201 High Street, FREMANTLE WA 6160  
Tel: 08 9335 2533 Fax: 08 9335 2544 e-mail: [training@iea.com.au](mailto:training@iea.com.au)