

## Voyage of the Web Crawler

Ian Burke  
11/20/2006

I don't know where to find that? They say you can find everything on the Internet but I just don't seem to be able to find anything. That is so often how people feel when they find themselves looking out onto the World Wide Web. But there are a few tricks that can go a long way to helping to tame the web.

Forge ahead noble voyager. Know no fear as you journey across the World Wide Web. You may have no doubt or trepidation as you download files to your laptop. Whether you are a timid or a fearless Internet user there are things to know and safe practices to use as you browse the Web.

Anti-virus software is a must have regardless of whether it is free software from nonprofit Internet consortium ClamAV, or Symantec Corporation's commercial package. If you are going to spend any time on the Internet you are going to get a virus and you need to be protected. The same is true about adware software. With the first pop-up add that shows up on your browser you can be fairly certain that you have much more behind the scenes on your computer. Good adware software such as Ad-aware or Spybot will help you find little problem files that are the cause of all this fuss.

But, that is the easy part, putting protection on your computer. The hard part is knowing what to do on the Net: knowing how to find stuff and knowing where to and where not to go. Rule number one, if you don't know it don't go there. It is sort of like the rules our parents gave us when we were kids, "Never go to a stranger's house or car." Well it is still true. That is not to say you can not go to a new web site, but you want to know something about it first. Here are ten questions that can help you decide if the site is safe. It is not saying that you need to answer them all, they are just guidelines.

- Who owns the site?
  - Do they put their name right on the front of the page? It should be in the top of your web browser (the title bar for [www.microsoft.com](http://www.microsoft.com) reads, "Microsoft Corporation".)
- How did I get to the site?
  - If you followed a link where was that link; was it in your email, another web page, or in a document? Each one of these or other sources has a different level of legitimacy.
- Who else references or links to this site?
  - Is this a site that a lot of people refer to such as Amazon.com or is it a page that you found in a random search?
- What is the site about?
  - The site content is invaluable. If the content is legitimate, scholarly, and of a reputable source (i.e. a university or known organization) there is less of a chance that the author is going to be out to get you.
- Does the site have downloadable content?
  - A text site is going to have a harder time hurting you than when you download something.
- Does the site have pop-ups or place cookies on my system?

- Cookies and ActiveX controls are small files that your web browser uses to customize how the web page looks for you. An attacker can also use them to infect your system or to gather information about your system.
- Who do I know that has been to this site before?
  - Site references are always good.
- How long has this site been around?
  - Sometimes well established sites become targets for the bad guys but generally the Googles, IBMs, and Yahoos of the Web are fairly safe.
- Who is the site owner affiliated with?
  - If you are looking at a site and on their header they identify themselves as a division of Microsoft, such as Office.Live.Com, chances are fairly good that the site is as safe as a Microsoft site.
- Who hosts the site? (this may be different than the owner)
  - This is not always easy to figure out but the party hosting the site is the one that usually has all of the virus, spam, adware and other filters on the site. If you go to my personal security forum ianburke.net, you are actually going to a Microsoft, Office.Live.Com site hosted by Microsoft, fairly secure.

After you have answered a few of these questions you may want to look at your tools. Know your search engine. Yahoo and MSN are nice in that their search spectrum is fairly narrow. Google actually researches the sites that they add to their crawler. That does not imply that only sites they have researched will show up in a Google search, but rather helps to improve the accuracy of their results and improves the number of links in their searches. DogPile combines several search engines together. With each of these you should know how to conduct your searches. Google Help: <http://www.google.com/help/features.html#link> explains the many different ways that you can conduct a search through Google. It helps to narrow down your search so that you are not wading through garbage. When you are looking at more legitimate data you are less likely to run into those sites that have the viruses and malware.

So when do you download? There are three absolutes:

- You must know the site personally.
- You must know the file. Usually there are check files to verify for this reason.
- You must be able to certify the download site.

Just because they said it is a Microsoft file does not mean it is a Microsoft site. Every Microsoft download will offer a certificate that you can view. Every Sourceforge.com download will come with a check sum (a hashed mathematical value) that you can verify and a mirror site you can use if you are not comfortable with the site you are at. Before you download something from the web verify the file and verify the source, then do it again.

The World Wide web can be a great tool and a lot of fun. Preparing yourself for your adventure and using safe practices while you are surfing can make the grand adventure a rewarding one.