CHAPTER 1

An introduction to the internet

Introduction
This is the fourth edition of the Advanced Internet Searcher’s Handbook, and in the three previous editions I felt that it was necessary to include an introduction to the internet. I explained exactly what it was and how it worked in general terms. Looking back at the section headings it all seemed to be a wonderfully innocent time – ‘both local and global’, ‘it isn’t a single entity’, ‘it’s difficult to say who is in charge’, and so on. I was tempted to write a totally different introduction to this edition, but in actual fact the subheadings that I used are still in many cases apposite today. The internet is, if anything, even more local and global, and it’s certainly not a single entity. Internet search is still very much centre stage, and while we have had (and to a limited extent still have) browser wars, a key ‘battleground’ is still being fought over search.

This may come as something of a surprise to anyone who thinks that Google has routed the opposition and sent them fleeing. However, this is most certainly not the case. While there’s no doubt that Google is – at the time of writing – the dominant force when it comes to search, there is no saying that it will always stay there. Microsoft, Facebook, and to a lesser extent perhaps Apple and Amazon, are all keen on increasing their own market share. The war continues, but instead of the basic concept of ‘largest search engine wins’ the battles are now being fought across social media, in networks, on laptops and more importantly, on mobile phones. Internet search is not all about Google, and although I’ll be talking about Google a lot in this edition (in contrast to the first edition, when it wasn’t mentioned at all), I’ll also be referring to the dozens of other search engines that are out there, and they also have an important part to play in the development of internet search into the next decade.
Consequently, I’ll stay with tradition, to an extent at least, and have as my opening chapter a discussion of the internet, although my main focus will be on the way that search has infiltrated almost every area of the time we spend on the internet. I’ll then start to look at search in more depth.

**An overview of the internet**

The internet is almost, but not quite, ubiquitous. The UK Office of National Statistics, in their 2011 release ‘Internet Access, Households and Individuals’ (www.ons.gov.uk/ons/rel/rdit2/internet-access---households-and-individuals/2011/stb-internet-access-2011.html) reported that 77% of UK households had internet access, and 45% of users have used a mobile phone to connect to it. People use the internet to communicate with friends either by e-mail or increasingly via a social network (or two!), purchase products, order their shopping, chat real-time face to face via video chat, complete official forms, obtain information from local authorities, consume news, watch television or films, or ensure the safety of their house while on a holiday that was booked and paid for via their online bank.

The two things that most people will do while using the internet however are e-mail and search. These activities have consistently been shown to be the most important things that we do on the net. Statistics from a Pew Internet survey from May 2011 show that both of these trend at about 90% or more and have done so since 2005 (www.pewinternet.org/Reports/2011/Search-and-email/Report.aspx).

We search for information all the time – and thanks to Google’s web and trends history, which keeps tabs on me, I know that on average I run 1470 searches a month, which is about 48 per day. I search most on a Wednesday, least on a Sunday and my busiest search time of day is between 11 a.m. and 1 p.m. I suspect that these figures are not that untypical, and in fact may well be below average if I limit my sample universe to library and information professionals doing their jobs. Of course, that figure only takes into account the searching that I do via Google; there are the dozens of other search engines that I use, the casual searching that I will do once I arrive at a site using their own search engines, or the time that I spend on Facebook or using my iPad. Further interrogation of the figures previously mentioned...
also show that internet search activity is not limited to particular demographic types – irrespective of race, ethnicity, age, education, income, everyone searches. To that extent, the internet has turned us all into researchers. However, due to the complexity of search and its paradoxical apparent ease, some people are rather better at it than others.

**What the internet is and is not**

**It’s not a single network**

In previous editions of the book I took time to explain how the internet was created and maintained, and was actually created from the interlinking of different networks. However, the point that’s worth making now, with respect to internet searching, is that data is still spread out into discrete units, hidden in databases that are only accessible via password, on social networks in walled gardens, kept private on closed wikis, found in social bookmarking systems, on photograph- or video-sharing websites, and so on. No single search engine can access all of this information and make it available, and in fact the hidden web is estimated to be thousands of times larger than the web content found by general search engines (http://websearch.about.com/od/invisibleweb/f/What-Is-The-Size-Of-The-Hidden-Web.htm). Consequently, rather than search becoming easier to do, it’s actually much harder. The cry often heard is ‘it’s all on Google, why do we need libraries or librarians?’, when in actual fact it’s hardly on Google at all, and we are in greater need of libraries and librarians to help us find the exact piece of information that we need. There’s a quote on the net which sums this situation up perfectly, attributed to the author Neil Gaiman, and goes like this: ‘Google can bring you back 100,000 answers, a librarian can bring you back the right one.’

Despite what the large search engines would have us believe, search has not become easier over the course of time – it’s become much more complex and will increasingly become so as more and more data floods onto internet-based servers and in ‘the cloud’.

**It’s both local and global**

In ‘the old days’ (and how glibly that phrase rolls off the tongue) I meant that information was ubiquitous and it didn’t matter where you were; if you had an internet connection you could get the information that you needed. I was of course referring to desktop use, but this is now an even more
accurate statement than it was in the past. If I have my smartphone with me, not only can I access it to get the information that I need, I can simply speak to the search engine (as long as I’m using a Google app); it will know where in the world I am via a ‘discussion’ with the smartphone and it will bring me back information that I need based on where I am. If I’m in Manchester and I want hotels, that’s what I’ll get. If I’m in Liverpool and I need to find an excellent example of a football ground Google will happily direct me to Goodison Park, offering directions along the way. I can search for what I need by using Google Maps (http://maps.google.com), or I can seek out the information from a localized search engine, or check via Twitter’s geographic search function. Not only is it now possible to go to anywhere in the world to get the most appropriate information that I need, but the most appropriate local information comes to me.

It isn’t a single entity
When I first wrote this I was thinking about ‘the world wide web, newsgroups, mailing lists’ and more besides. During the course of the last decade newsgroups and mailing lists became less important and the rise of the website became the pre-eminent focus for most people. The web was the internet, and the internet was the web. However, even back in the early years of the 2000s it was possible to discern the beginnings of competition to the HTML page, in the form of weblogs, where people were able to create their own spaces to say whatever they wanted, without having to author websites and buy domain names. These days, as well as newsgroups and mailing lists (which do still exist, although their popularity has dwindled) we have blogs, wikis, social network resources, photograph-sharing websites, Twitter streams, postings to Google+, social bookmarking databases, collections of books and forums on booksharing websites such as LibraryThing. We have sites that store and make available user-created videos, slideshow presentations, podcasts and more. If anything, the internet is more disparate in 2011 than it was in 2001 and search has evolved, and is evolving, to cope with that. With respect to search, the key thing to keep in mind therefore is that no one single search engine can do the job. If you wish to search the internet well then it is necessary to have a good grasp and
understanding of a wide variety of different search engines and an awareness of their strengths and weaknesses.

**It is possible to use a wide variety of hardware and software**
Yet again, the accuracy of this statement has not changed, although both the content and concept have. As previously illustrated, an increasing number of people are using their smartphones to connect to the internet. However, it’s also possible to connect via games consoles, televisions and many other devices, to say nothing of the traditional laptop or desktop. Finally, we must not forget the access that we can get to the internet via various tablet devices. Since we can use a variety of different types of software and ‘the cloud’ to store our data we must once again remember that no single search engine is capable of searching all of these different resources. As the amount of hardware and software options increase, so too do the places in which to store and find information, and consequently search engines have to become much more sophisticated and provide a much wider variety of options to allow us to find the information we need.

Moreover, unlike the early days of the internet, when it was necessary to produce your information in a standardized format or a web page, this can now be achieved using a wide variety of different file formats. An effective search engine must be able to not only index data but provide you with various options to enable you to retrieve it effectively.

**It’s difficult to say who is in charge**
Publishing data onto the internet used to be a difficult task; it was necessary to understand the basics of HTML (HyperText Markup Language), then to obtain a website address, and to understand, or employ somebody who could understand, how to get a good ranking from search engines. One could, however, argue that even though no one individual or group was in charge of the internet various organizations could be in charge of discrete units of data. For example, a council, through the use of its website, could decide what information to make available, and if somebody wished to contact the council or to refer to the material that was available on the site...
or to discuss the services that were available, they would have to do it through the website. This no longer holds true, however, since with the increase in social media, and the ease with which it is now possible to publish your own data, it is now easier than ever for anyone to say whatever they would like to. If somebody wished to complain about council services they could do so easily via their own blog, on Twitter, or in a Facebook group. The importance of this cannot be overestimated, because the internet has now evolved from a resource where few published to the many into one where the many can publish to the many or to the few. As a result, the value of the website has decreased over time as the value of publishing via social media applications has increased. Not only has this meant that traditional search engines have had to adapt and to attempt to index that flood of data, it has also caused entirely new search engines to spring into existence to focus specifically on the information. Once again, therefore, if you wish to search the internet quickly and effectively, increasingly a single search engine will not do the job for you. Moreover, we have now moved to a situation where search is becoming real-time. By that I mean we are able to find information that was published not months or weeks or days ago but in the last few seconds. Consequently, our expectations of search engines have by necessity changed, for we not only expect a search engine to provide us with content from a wide variety of different types of resources but we also expect it to provide us with that data even if was only published a few seconds ago.

It’s fast and effective
As previously mentioned, publishing information on the internet can now be done in seconds and retrieved as quickly. When searching for information we now not only need to consider the authority of the data, its accuracy and the context within which it was made available, but we increasingly need to ask ourselves ‘how fresh does this information need to be?’ That should also affect the choice of search engine that we use but it also brings into sharp focus that the information is even less trustworthy now than it has ever been in the past, because we are all publishers, journalists and authors. A good search engine must therefore not only retrieve information for us from a wide variety of types of data but it also needs to do so quickly and effectively, and have some way of checking the authority of that information. It is no longer acceptable for a search engine to return results to us based on simple algorithms of search terms in the title
repetition, links, and so on. We need much more by the way of authority – and here search engines face a dilemma. If the traditional website becomes less important, what takes its place? Given the rise of social media and the ability of people to produce their own content, we must increasingly look towards the authority of individuals, and the ways in which that can be assessed, rather than that of the traditional website.

It’s easy to talk to individuals or groups
In previous editions of the book I used this section to explain that it was perfectly possible for people to use resources such as newsgroups and mailing lists to share information backwards and forwards. That ability has now increased exponentially as publishing on the web has become easier. One only has to look at the popularity of resources such as Twitter and Facebook to see that people enjoy communicating with each other, and communicating a lot! That communication can also take the form of documents being placed on the net or shared via social networks or stored in resources such as Flickr. The difficulty of searching for and finding information is therefore exacerbated by the ease and availability of mechanisms for producing it.

It’s not all hard work
I used to take this to mean that there were plenty of resources available on the internet for people to indulge their hobbies. I think these days we can take that as a given, but if we start to look at this in terms of internet search and the effect that has we can very quickly see that people who have an interest in a particular subject or hobby are able to produce content and make it available on the web. Therefore, an increasing amount of authoritative content may well be produced by hobbyists or skilled amateurs and they will be making this information available in a variety of different ways on a variety of different platforms. Effective search engines must therefore once again be able to work out the authority and validity of content not by the relatively simple means of assessing a website but by the far more difficult method of evaluating the information provided by specific individuals. So we can see that, once again, search is becoming inextricably bound up with social media resources and if a search engine is going to succeed it has to be able to take individuals into account at least as much as, if not more than, traditional websites.
It’s not just for ‘geeks’

It used to be quite complicated to put information onto the internet, and there once was a concern that it would become more difficult, with complex HTML coding and markup. However, the exact opposite has occurred and it has become far easier to share content, photographs and images. Moreover, with the advent of smartphones people are able to add content to the internet quickly and effectively without even thinking about it. Not only are we all journalists, we are instant reporters and able to provide information on newsworthy items as they are occurring. The influence that this has on the development of search engines is therefore quite profound, since they have to weigh up the various merits of the immediacy of the data with the uncertain authority of the person producing it.

If we look at this subject from another angle we also have to take into account the fact that if anybody can produce content and make it available on the web anyone and everybody is going to want to search it. Search engines therefore are caught in a cleft stick, because on the one hand they want to make their engine as quick, as simple and as easy to use as possible for an unskilled and technically illiterate user, but on the other they have to be able to make it extraordinarily sophisticated, so that we can get exactly the information that we need when we want it. For search engines this is a major issue and the ‘Holy Grail’, because they want to personalize their results in order to make them appropriate for the needs and requirements of each individual. The danger is that they will create what Eli Pariser has termed ‘filter bubbles’ (www.ted.com/talks/eli_pariser_beware_online_filter_bubbles.html); by this he means that a search engine will give us the results that it thinks we want, without exposing us to different or alternative viewpoints.

A secondary consideration here is that the more search engines personalize results, the less a generic concept of search can exist. We are already seeing this, in effect, as an increasing number of companies will suggest in their advertising that we search for a specific word or phrase. It’s also worth pointing out that companies will often point users towards their Facebook pages rather than their websites. As we shall see, the rise of Facebook and its ‘walled garden’ is a direct threat to traditional search engines, which emphasize their role in finding information on the web generally, on web pages in websites. If users are going to spend their time in Facebook (as they do) and run searches using Facebook’s search option, why would they want to leave that environment to search elsewhere? As a result, it becomes more and more difficult to respond to an enquirer by
giving them information couched in the terms ‘do a search for so and so, and you will find that the result that you need is the third one down’, since the third result down for the enquirer is not necessarily the same third result as seen by the searcher. There are, of course, many ways around this particular problem, such as using search engines which do not personalize results or logging out from those which do and then searching, thus ensuring that the engine doesn’t know who you are, and cannot use that knowledge to affect the returned results. It is all too easy to forget this step, but skilled searchers will always keep in the back of their mind the fact that they are often searching on behalf of other people rather than themselves, and so the personalized results which they may receive are not necessarily the best results for the original enquirer.

It’s not well organized
I don’t think there can be any doubt about this, and indeed I would suggest that it is becoming less and less organized every second. This again means that search engines and well honed search techniques become absolutely vital. The more often people say ‘it’s all on Google’, the more experienced searchers have to point out the dangerous fallacy in that statement.

It’s growing at an enormous rate
Little did I realize when I first wrote those words just how quickly data on the internet is increasing. Twitter receives over 400 million tweets per day, Flickr users upload 3500 images per minute (www.digitalbuzzblog.com/infographic-the-growth-of-social-media-2011). There are 152 million blogs on the internet, 40 hours’ worth of videos are uploaded to YouTube every minute and on English Wikipedia over 30,000 editors make more than 5 edits per month (www.bbc.co.uk/news/magazine-18892510). Around 1.4 million blog posts are created every day and 3.5 billion pieces of content are shared every week on Facebook (http://visual.ly/internet-full). It is estimated that by 2014 users will create and replicate 5 zettabytes worth of data (1 zettabyte is 1 trillion gigabytes) and that by 2020 this figure will have risen to 35 zettabytes (www.economist.com/node/21553410?fsrc=dg|b). While these figures are almost impossible to comprehend, one thing is for sure – by the time you read them they are going to be woefully inaccurate, such is the speed of growth of internet content.
Consequently, it is simply not possible for one search engine, or indeed one dozen search engines, to index, arrange and make available that amount of information. Unfortunately, however, it is the job of skilled information researchers to find as much appropriate information as they can out of that morass of data and to package it and make it available for their enquirers. So, rather than making things simpler and easier, the easy access to applications which make us all creators of content has in actual fact made it that much harder to find the information that we require. In order to solve this conundrum search engines have to improve, people have to help other people, and searchers need to explore as many avenues as they possibly can in order to do their job effectively.

Search engines
I have already broadly discussed some of the changes that have occurred with search engines and search engine technology in the last few years, but at this point it is worthwhile spending a few moments talking about another very important aspect of search engines, which is the understandable desire of their owners to create as much revenue as they can. So let me be clear right from the outset: the vast majority of search engines do not have at their core an interest in helping you find information – they exist in order to display as much advertising on your screen as possible. If there was no money to be made from advertising, there would be no Google and there would be no Facebook. A second and equally cynical point also needs to be made, which is that you are not a user of search engines; the users of search engines are the advertisers, and all of us are simply the fodder that is required in order to make them money.

Search engines do not have to be excellent at search; they merely have to be good enough for the vast majority of users. In other words, search engine developers will only introduce new functionality and techniques if and when they think it will increase their revenue streams. In the autumn of 2011 Google dramatically closed down a large number of experiments that they were running in the Google labs. Many of these experiments were interesting from a search standpoint but would have had little interest for the majority of users. Since Google wishes to develop a social network to rival that of Facebook many of their programmers were moved off projects that were interesting but were not likely to make the company much money, and onto the development of Google+, which is a central plank in the company’s development of their revenue streams.
What does this mean for the searcher? First, it means that you should have no loyalty whatsoever to any search engine at all. Never forget that you are not a user of your favourite search engine, you are its cattle. The idea of brand loyalty with respect to search engines should simply not exist. You should exhibit the behaviour of a discerning shopper in a large mall; go from counter to counter to seek out the best bargains – or from engine to engine to seek out the best functionality that you can – and then move on to the next. Secondly, always remember that new search engines arrive on the scene (I am generally told about one new search engine per day, which is just slightly more than the number that close) and they should be explored to see what they can offer you as the searcher. Be ruthless in your approach and once a search engine fails to develop or give you what you need quickly and effectively, drop it and move on. Do not get attached to a search engine; it is in its interests to get you to go back to it over and over again so that it can show you more adverts, while it is in your interest to find the best resource for the query that you have. Once you start to rely on a specific search engine you will become a less successful searcher. That does not of course mean that you can’t have your preferences, because we all do, but it does mean that your head needs to rule your heart.

Having said all of that, I really don’t want to paint a bleak picture of internet search engines or to be overly cynical. Internet search is a vibrant and exciting area and one that is constantly changing. For the last 15 years I have run a course entitled ‘advanced internet searching’ and it is still as popular today as it has ever been. While the concept of the course is the same as it always has been, the content changes from month to month as I find new search engines or existing search engines develop new functionality. As we shall see in later chapters, there are many opportunities to explore different resources many of which are far more powerful than we would have thought possible even three or four years ago.

Failing resources
However, it should also be mentioned that many resources have declined in use and popularity as other resources have increased in value. There is no point in devoting entire chapters to resources such as USENET newsgroups, intelligent agents and mailing lists; these days they are worth little more than footnotes. The reasons for this are many and varied. While USENET newsgroups still exist, they are increasingly underutilized, because in order to get the most out of them it is necessary to use...
specialized software. Moreover, since they are almost entirely unregulated, they have fallen victim to the spammers and while several people made valiant attempts to save them they are quickly passing into internet history and have largely been replaced by resources such as Facebook groups.

Intelligent agents first came to the fore in the early years of this century, and they were supposed to hunt around on the internet and find the information that we needed and bring it back to us. The term has since fallen into disuse, but the idea of news collation and curation on the other hand has developed and has been incorporated into other resources. Many readers will be familiar with tools such as Zite and Flipboard that they use on their tablets and these are the descendants of intelligent agents.

Mailing lists continue to exist, although I think that it is fair to say that they are in the main within the remit of academia rather than general use. They flourished in the past due to their use of e-mail, as it was very easy for people to disseminate the messages that they wished to larger groups of people and to have general discussions. However, with the increase in blogs, Twitter, Facebook groups and other social networks those discussions have generally migrated away from e-mail towards social media. We cannot and should not forget the existence of these resources, however, because, depending on the enquiry that we have, their archival value is still very important. The data on websites tends in the main to get overwritten but the archives of USENET and some mailing lists are still safely stored away.

**The information mix**

Effective internet searching is part science, part art, part skill and part luck. Most of all, however, it is an effective blend of different resources, different search engines and a willingness to look for information wherever it might be and for however long it takes. None of that should come as any sort of surprise to an information professional but the average internet user will probably not have the experience, knowledge, understanding or tenacity to blend those together in a successful search. Of course, in this book I am going to spend very little time talking about resources that are not on the internet, but we should never forget that traditional resources such as books, databases on CD-ROM, or even microfiche still have an important part to play. Therefore, please do consider their value and encourage your users to explore them because, after all, a successful search is one where you get the right answer, irrespective of where it comes from.
Summary

In this opening chapter I hope that I have at least in part laid down some markers that relate to where internet search is currently and where I believe that it will move to within the next few years. While it is all too easy to agree with the statement ‘Everything is on Google’, it is not only a fallacy but a dangerous one at that; I hope that this will not come as a surprise to anyone reading this book. In the coming chapters I will explore the different types of search engine that are available and their advantages and disadvantages. I will also provide examples of search engines within each category (full-text, directory, multi-search, and so on) and look at ways in which you can get the most out of them. Internet searching is a fascinating and enjoyable activity – in some cases almost a hobby – and I hope that you will be as excited, enthralled and intrigued as I am when I get the opportunity to explore a new search engine for the first time.