Chapter 1

The Internet and the Visible Web

To understand the Web in the broadest and deepest sense, to fully partake of the vision that I and my colleagues share, one must understand how the Web came to be.

—Tim Berners-Lee, *Weaving the Web*

Most people tend to use the words "Internet" and "Web" interchangeably, but they're not synonyms. The Internet is a networking protocol (set of rules) that allows computers of all types to connect to and communicate with other computers on the Internet. The Internet's origins trace back to a project sponsored by the U.S. Defense Advanced Research Agency (DARPA) in 1969 as a means for researchers and defense contractors to share information (Kahn, 2000).

The World Wide Web (Web), on the other hand, is a software protocol that runs on top of the Internet, allowing users to easily access files stored on Internet computers. The Web was created in 1990 by Tim Berners-Lee, a computer programmer working for the European Organization for Nuclear Research (CERN). Prior to the Web, accessing files on the Internet was a challenging task, requiring specialized knowledge and skills. The Web made it easy to retrieve a wide variety of files, including text, images, audio, and video by the simple mechanism of clicking a hypertext link.

DEFINITION

Hypertext

A system that allows computerized objects (text, images, sounds, etc.) to be linked together. A hypertext link points to a specific object, or a specific place with a text; clicking the link opens the file associated with the object.

The primary focus of this book is on the Web—and more specifically, the parts of the Web that search engines can't see. To fully understand the phenomenon called the Invisible Web, it's important to first understand the fundamental differences between the Internet and the Web.

In this chapter, we'll trace the development of some of the early Internet search tools, and show how their limitations ultimately spurred the popular acceptance of the Web. This historical background, while fascinating in its own right, lays the foundation for understanding why the Invisible Web could arise in the first place.

How the Internet Came to Be

Up until the mid-1960s, most computers were stand-alone machines that did not connect to or communicate with other computers. In 1962 J.C.R. Licklider, a professor at MIT, wrote a paper envisioning a globally connected "Galactic Network" of computers (Leiner, 2000). The idea was far-out at the time, but it caught the attention of Larry Roberts, a project manager at the U.S. Defense Department's Advanced Research Projects Agency (ARPA). In 1966 Roberts submitted a proposal to ARPA that would allow the agency's numerous and disparate computers to be connected in a network similar to Licklider's Galactic Network.

Roberts' proposal was accepted, and work began on the "ARPANET," which would in time become what we know as today's Internet. The first "node" on the ARPANET was installed at UCLA in 1969 and gradually, throughout the 1970s, universities and defense contractors working on ARPA projects began to connect to the ARPANET.

In 1973 the U.S. Defense Advanced Research Projects Agency (DARPA) initiated another research program to allow networked computers to communicate transparently across multiple linked networks. Whereas the ARPANET was just one network, the new project was designed to be a "network of networks." According to Vint Cerf, widely regarded as one of the "fathers" of the Internet, "This was called the Internetting project and the system of networks which emerged from the research was known as the 'Internet'" (Cerf, 2000).

It wasn't until the mid 1980s, with the simultaneous explosion in use of personal computers, and the widespread adoption of a universal standard of Internet communication called Transmission Control Protocol/Internet Protocol (TCP/IP), that the Internet became widely available to anyone desiring to connect to it. Other government agencies fostered the growth of the Internet by contributing communications "backbones" that were specifically designed to carry Internet traffic. By the late 1980s, the Internet had grown from its initial network of a few computers to a robust communications network supported by governments and commercial enterprises around the world.

Despite this increased accessibility, the Internet was still primarily a tool for academics and government contractors well into the early 1990s. As more and more computers connected to the Internet, users began to demand tools that would allow them to search for and locate text and other files on computers anywhere on the Net.

Early Net Search Tools

Although sophisticated search and information retrieval techniques date back to the late 1950s and early '60s, these techniques were used primarily in closed or proprietary systems. Early Internet search and retrieval tools lacked even the most basic capabilities, primarily because it was thought that traditional information retrieval techniques would not work well on an open, unstructured information universe like the Internet.

Accessing a file on the Internet was a two-part process. First, you needed to establish direct connection to the remote computer where the file was located using a terminal emulation program called Telnet. Then you needed to use another program, called a File Transfer Protocol (FTP) client, to fetch the file itself. For many years, to access a file it was necessary to know both the address of the computer and the exact location and name of the file you were looking for—there were no search engines or other file-finding tools like the ones we're familiar with today.

DEFINITIONS

File Transfer Protocol (FTP)
A set of rules for sending and receiving files of all types between computers connected to the Internet.

Telnet
A terminal emulation program that runs on your computer, allowing you to access a remote computer via a TCP/IP network and execute commands on that computer as if you were directly connected to it. Many libraries offered telnet access to their catalogs.

Thus, "search" often meant sending a request for help to an e-mail message list or discussion forum and hoping some kind soul would respond with the details you needed to fetch the file you were looking for. The situation improved somewhat with the introduction of "anonymous" FTP servers, which were centralized file-servers specifically intended for enabling the easy sharing of files. The servers were anonymous because they were not password protected—anyone could simply log on and request any file on the system.

Files on FTP servers were organized in hierarchical directories, much like files are organized in hierarchical folders on personal computer systems today. The hierarchical structure made it easy for the FTP server to display a directory listing of all the files stored on the server, but you still needed good knowledge of the contents of the FTP server. If the file you were looking for didn't exist on the FTP server you were logged into, you were out of luck.

The first true search tool for files stored on FTP servers was called Archie, created in 1990 by a small team of systems administrators and graduate students at McGill University in Montreal. Archie was the prototype of today's search engines, but it was primitive and extremely limited compared to what we have today. Archie roamed the Internet searching for files available on anonymous FTP servers, downloading directory listings of every anonymous FTP server it could find. These listings were stored in a central, searchable database called the Internet Archives Database at McGill University, and were updated monthly.

Although it represented a major step forward, the Archie database was still extremely primitive, limiting searches to a specific file name, or for computer programs that performed specific functions. Nonetheless, it proved extremely popular—nearly 50 percent of Internet traffic to Montreal in the early '90s was Archie related, according to Peter Deutsch, who headed up the McGill University Archie team.

"In the brief period following the release of Archie, there was an explosion of Internet-based research projects, including WWW, Gopher, WAIS, and others" (Deutsch, 2000).

"Each explored a different area of the Internet information problem space, and each offered its own insights into how to build and deploy Internet-based services," wrote Deutsch. The team licensed Archie to others, with the first shadow sites launched in Australia and Finland in 1992. The Archie network reached a peak of 63 installations around the world by 1995.

Gopher, an alternative to Archie, was created by Mark McCahill and his team at the University of Minnesota in 1991 and was named for the university's mascot, the Golden Gopher. Gopher essentially combined the Telnet and FTP protocols, allowing users to click hyperlinked menus to access information on demand without resorting to additional commands. Using a series of menus that allowed the user to drill down through successively more specific categories, users could ultimately access the full text of documents, graphics, and even music files, though not integrated in a single format. Gopher made it easy to browse for information on the Internet.

According to Gopher creator McCahill, "Before Gopher there wasn't an easy way of having the sort of big distributed system where there were seamless pointers between stuff on one machine and another machine. You had to know the name of this machine and if you wanted to go over here you had to know its name.

"Gopher takes care of all that stuff for you. So navigating around Gopher is easy. It's point and click typically. So it's something that anybody could use to find things. It's also very easy to put information up so a lot of people started running servers themselves and it was the first of the easy-to-use, no muss, no fuss, you can just crawl around and look for information tools. It was the one that wasn't written for techies."

Gopher's "no muss, no fuss" interface was an early precursor of what later evolved into popular Web directories like Yahoo!. "Typically you set this up so that you can start out with [a] sort of overview or general structure of a bunch of information, choose the items that you're interested in to move into a more specialized area and then either look at items by browsing around and finding some documents or submitting searches," said McCahill.

A problem with Gopher was that it was designed to provide a listing of files available on computers in a specific location—the University of Minnesota, for example. While Gopher servers were searchable, there was no centralized directory for searching all other computers that were both using Gopher and connected to the Internet, or "Gopherspace" as it was called. In November 1992, Fred Barrie and Steven Foster of the University of Nevada System Computing Services group solved this problem, creating a program called Veronica, a centralized Archie-like search tool for Gopher files. In 1993 another program called Jughead added keyword search and Boolean operator capabilities to Gopher search.

DEFINITIONS

Keyword

A word or phrase entered in a query form that a search system attempts to match in text documents in its database.

Boolean

A system of logical operators (AND, OR, NOT) that allows true-false operations to be performed on search queries, potentially narrowing or expanding results when used with keywords.

Popular legend has it that Archie, Veronica and Jughead were named after cartoon characters. Archie in fact is shorthand for "Archives." Veronica was likely named after the cartoon character (she was Archie's girlfriend), though it's officially an acronym for "Very Easy Rodent-Oriented Net-Wide Index to Computerized Archives." And Jughead (Archie and Veronica's cartoon pal) is an acronym for "Jonzy's Universal Gopher Hierarchy Excavation and Display," after its creator, Rhett "Jonzy" Jones, who developed the program while at the University of Utah Computer Center.

Myth: The Web and the Internet Are the Same

The Internet is the world's largest computer network, made up of millions of computers. It's really nothing more than the "plumbing" that allows information of various kinds to flow from computer to computer around the world.

The Web is one of many interfaces to the Internet, making it easy to retrieve text, pictures, and multimedia files from computers without having to know complicated commands. You just click a link and voila: You miraculously see a page displayed on your browser screen.

The Web is only about 10 years old, whereas the Internet is 30-something. Prior to the Web, computers could communicate with one another, but the interfaces weren't as slick or easy to use as the Web. Many of these older interfaces, or "protocols," are still around and offer many different, unique ways of communicating with other computers (and other people).

Other Internet protocols and interfaces include:

- E-mail
- Forums & Bulletin Boards
- Internet Mailing Lists
- Newsgroups
- Peer-to-Peer file sharing systems, such as Napster and Gnutella
- Databases accessed via Web interfaces

As you see, the Internet is much more than the Web. In fact, the last item on the list above, databases accessed via Web interfaces, make up a significant portion of the Invisible Web. Later chapters will delve deeply into the fascinating and tremendously useful world of Web accessible databases.

A third major search protocol developed around this time was Wide Area Information Servers (WAIS). Developed by Brewster Kahle and his colleagues at Thinking Machines, WAIS worked much like today's metasearch engines. The WAIS client resided on your local machine, and allowed you to search for information on other Internet servers using natural language, rather than using computer commands. The servers themselves were responsible for interpreting the query and returning appropriate results, freeing the user from the necessity of learning the specific query language of each server.

WAIS used an extension to a standard protocol called Z39.50 that was in wide use at the time. In essence, WAIS provided a single computer-to-computer protocol for searching for information. This information could be text, pictures, voice, or formatted documents. The quality of the search results was a direct result of how effectively each server interpreted the WAIS query.

All of the early Internet search protocols represented a giant leap over the awkward access tools provided by Telnet and FTP. Nonetheless, they still dealt with information as discrete data objects. And these protocols lacked the ability to make connections between disparate types of information—text, sounds, images, and so on—to form the conceptual links that transformed raw data into useful information. Although search was becoming more sophisticated, information on the Internet lacked popular appeal. In the late 1980s, the Internet was still primarily a playground for scientists, academics, government agencies, and their contractors.

Fortunately, at about the same time, a software engineer in Switzerland was tinkering with a program that eventually gave rise to the World Wide Web. He called his program Enquire Within Upon Everything, borrowing the title from a book of Victorian advice that provided helpful information on everything from removing stains to investing money.

Enquire Within Upon Everything

> "Suppose all the information stored on computers everywhere were linked, I thought. Suppose I could program my computer to create a space in which anything could be linked to anything. All the bits of information in every computer at CERN, and on the planet, would be available to me

> and to anyone else. There would be a single, global information space.
>
> "Once a bit of information in that space was labeled with an address, I could tell my computer to get it. By being able to reference anything with equal ease, a computer could represent associations between things that might seem unrelated but somehow did, in fact, share a relationship. A Web of information would form."
>
> —Tim Berners-Lee, *Weaving the Web*

The Web was created in 1990 by Tim Berners-Lee, who at the time was a contract programmer at the Organization for Nuclear Research (CERN) high-energy physics laboratory in Geneva, Switzerland. The Web was a side project Berners-Lee took on to help him keep track of the mind-boggling diversity of people, computers, research equipment, and other resources that are de rigueur at a massive research institution like CERN. One of the primary challenges faced by CERN scientists was the very diversity that gave it strength. The lab hosted thousands of researchers every year, arriving from countries all over the world, each speaking different languages and working with unique computing systems. And since high-energy physics research projects tend to spawn huge amounts of experimental data, a program that could simplify access to information and foster collaboration was something of a Holy Grail.

Berners-Lee had been tinkering with programs that allowed relatively easy, decentralized linking capabilities for nearly a decade before he created the Web. He had been influenced by the work of Vannevar Bush, who served as Director of the Office of Scientific Research and Development during World War II. In a landmark paper called "As We May Think," Bush proposed a system he called MEMEX, "a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility" (Bush, 1945).

The materials stored in the MEMEX would be indexed, of course, but Bush aspired to go beyond simple search and retrieval. The MEMEX would allow the user to build conceptual "trails" as he moved from document to document, creating lasting associations between different components of the MEMEX that could be recalled at a later time. Bush called this "associative indexing … the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the MEMEX. The process of tying two items together is the important thing."

In Bush's visionary writings, it's easy for us to see the seeds of what we now call hypertext. But it wasn't until 1965 that Ted Nelson actually described a computerized system that would operate in a manner similar to what Bush envisioned. Nelson called his system "hypertext" and described the next-generation MEMEX in a system he called Xanadu.

Nelson's project never achieved enough momentum to have a significant impact on the world. Another twenty years would pass before Xerox implemented the first mainstream hypertext program, called NoteCards, in 1985. A year later, Owl Ltd. created a program called Guide, which functioned in many respects like a contemporary Web browser, but lacked Internet connectivity.

Bill Atkinson, an Apple Computer programmer best known for creating MacPaint, the first bitmap painting program, created the first truly popular hypertext program in 1987. His HyperCard program was specifically for the Macintosh, and it also lacked Net connectivity. Nonetheless, the program proved popular, and the basic functionality and concepts of hypertext were assimilated by Microsoft, appearing first in standard help systems for Windows software.

Weaving the Web

The foundations and pieces necessary to build a system like the World Wide Web were in place well before Tim Berners-Lee began his tinkering. But unlike others before him, Berners-Lee's brilliant insight was that a simple form of hypertext, integrated with the universal communication protocols offered by the Internet, would create a platform-independent system with a uniform interface for any computer connected to the Internet. He tried to persuade many of the key players in the hypertext industry to adopt his ideas for connecting to the Net, but none were able to grasp his vision of simple, universal connectivity.

So Berners-Lee set out to do the job himself, creating a set of tools that collectively became the prototype for the World Wide Web (Connolly, 2000). In a remarkable burst of energy, Berners-Lee began work in October 1990 on the first Web client—the program that allowed the creation, editing, and browsing of hypertext pages. He called the client WorldWideWeb, after the mathematical term used to describe a collection of nodes and links in which any node can be linked to any other. “Friends at CERN gave me a hard time, saying it would never take off—especially since it yielded an acronym that was nine syllables long when spoken,” he wrote in *Weaving the Web*.

To make the client simple and platform independent, Berners-Lee created HTML, or HyperText Markup Language, which was a dramatically simplified version of a text formatting language called SGML (Standard Generalized Markup Language). All Web documents formatted with HTML tags would display identically on any computer in the world.

Next, he created the HyperText Transfer Protocol (HTTP), the set of rules that computers would use to communicate over the Internet and allow hypertext links to automatically retrieve documents regardless of their location. He also devised the Universal Resource Identifier, a standard way of giving documents on the Internet a unique address (what we call URLs today). Finally, he brought all of the pieces together in the form of a Web server, which stored HTML documents and served them to other computers making HTTP requests for documents with URLs.

Berners-Lee completed his work on the initial Web tools by Christmas 1990. In little more than two months he had created a system that had been envisioned for decades. Building the tools was the easy part. The hard part was to get the world to embrace the Web.

Berners-Lee wrote the original programs for the Web on the NeXT system, but thanks to his tireless efforts to persuade others to use the system, there were soon Web clients and servers available in a variety of different operating systems. As more and more people within CERN began to use the Web, the initial skepticism began to wear away.

Berners-Lee began actively to promote the Web outside of the lab, attending conferences and participating in Internet mailing and discussion lists. Slowly, the Web began to grow as more and more people implemented clients and servers around the world. There were really two seminal events that sparked the explosion in popular use of the Web. The first was the development of graphical Web browsers, including Voila, Mosaic, and others that integrated text and images into a single browser window. For the first time, Internet information could be displayed in a visually appealing format previously limited to CD-ROM-based multimedia systems. This set off a wave of creativity among Web users, establishing a new publishing medium that was freely available to anyone with Internet access and the basic skills required to design a Web page.

Then in 1995, the U.S. National Science Foundation ceased being the central manager of the core Internet communications backbone, and transferred both funds and control to the private sector. Companies were free to register "dot-com" domain names and establish an online presence. It didn't take long for business and industry to realize that the Web was a powerful new avenue for online commerce, triggering the dot-com gold rush of the late 1990s.

Early Web Navigation

The Web succeeded where other early systems failed to catch on largely because of its decentralized nature. Despite the fact that the first servers were at CERN, neither Berners-Lee nor the lab exercised control over who put up a new server anywhere on the Internet. Anyone could establish his or her own Web server. The only requirement was to link to other servers, and inform other Web users about the new server so they could in turn create links back to it.

But this decentralized nature also created a problem. Despite the ease with which users could navigate from server to server on the Web simply by clicking links, navigation was ironically becoming more difficult as the Web grew. No one was "in charge" of the Web; there was no central authority to create and maintain an index of the growing number of available documents. To facilitate communication and cross-linkage between early adopters of the Web, Berners-Lee established a list Web of servers that could be accessed via hyperlinks. This was the first Web directory. This early Web guide is still online, though most of the links are broken (http://www.w3.org/History/19921103-hypertext/hypertext/DataSources/bySubject/Overview. html).

Beyond the list of servers at CERN, there were few centralized directories, and no global Web search services. People notified the world about new Web pages in much the same way they had previously announced new Net resources, via e-mail lists or online discussions. Eventually, some enterprising observers of the Web began creating lists of links to their favorite sites. John Makulowich, Joel Jones, Justin Hall, and the people at O'Reilly & Associates publishing company were among the most noted authors maintaining popular link lists.

Eventually, many of these link lists started "What's New" or "What's Cool" pages, serving as de facto announcement services for new Web pages. But they relied on Web page authors to submit information, and the Web's relentless growth rate ultimately made it impossible to keep the lists either current or comprehensive.

What was needed was an automated approach to Web page discovery and indexing. The Web had now grown large enough that information scientists became interested in creating search services specifically for the Web. Sophisticated information retrieval techniques had been available since the early 1960s, but they were only effective when searching closed, relatively structured databases. The open, laissez-faire nature of the Web made it too messy to easily adapt traditional information retrieval techniques. New, Web-centric approaches were needed.

But how best to approach the problem? Web search would clearly have to be more sophisticated than a simple Archie-type service. But should these new "search engines" attempt to index the full text of Web documents, much as earlier Gopher tools had done, or simply broker requests to local Web search services on individual computers, following the WAIS model?

The First Search Engines

Tim Berners-Lee's vision of the Web was of an information space where data of all types could be freely accessed. But in the early days of the Web, the reality was that most of the Web consisted of simple HTML text documents. Since few servers offered local site search services, developers of the first Web search engines opted for the model of indexing the full text of pages stored on Web servers. To adapt traditional information retrieval techniques to Web search, they built huge databases that attempted to replicate the Web, searching over these relatively controlled, closed archives of pages rather than trying to search the Web itself in real time. With this fateful architectural decision, limiting search engines to HTML text documents and essentially ignoring all other types of data available via the Web, the Invisible Web was born.

The biggest challenge search engines faced was simply locating all of the pages on the Web. Since the Web lacked a centralized structure, the only way for a search engine to find Web pages to index was by following links to pages and gathering new links from those pages to add to the queue to visit for indexing. This was a task that required computer assistance, simply to keep up with all of the new pages being added to the Web each day.

But there was a subtler problem that needed solving. Search engines wanted to fetch and index all pages on the Web, but the search engines frequently revisited popular pages at the expense of new or obscure pages, because popular pages had the most links pointing to them—which the crawlers naturally followed. What was needed was an automated program that had a certain amount of intelligence, able to recognize when a link pointed to a previously indexed page and ignoring it in favor of finding new pages.

These programs became known as Web robots—“autonomous agents” that could find their way around the Web discovering new Web pages. Autonomous is simply a fancy way of saying that the agent programs can do things on their own without a person directly controlling them, and that they have some degree of intelligence, meaning they can make decisions and take action based on these decisions.

In June 1993 Mathew Gray, a physics student at MIT, created the first widely recognized Web robot, dubbed the “World Wide Web Wanderer.” Gray’s interest was limited to determining the size of the Web and tracking its continuing growth. The Wanderer simply visited Web pages and reported on their existence, but didn’t actually fetch or store pages in a database. Nonetheless, Gray’s robot led the way for more sophisticated programs that would both visit and fetch Web pages for storage and indexing in search engine databases.

The year 1994 was a watershed one for Web search engines. Brian Pinkerton, a graduate student in Computer Sciences at the University of Washington, created a robot called WebCrawler in January 1994. Pinkerton created his robot because his school friends were always sending him e-mails about the cool sites they had found on the Web, and Pinkerton didn’t have time to surf to find sites on his own—he wanted to “cut to the chase” by searching for them directly. WebCrawler went beyond Gray’s Wanderer by actually retrieving the full text of Web documents and storing them in a keyword-searchable database. Pinkerton made WebCrawler public in April 1994 via a Web interface. The database contained entries from about 6,000 different servers, and after a week was handling 100+ queries per day. The first Web search engine was born.

The image evoked by Pinkerton’s robot “crawling” the Web caught the imagination of programmers working on automatic indexing of the Web. Specialized search engine robots soon became known generically as "crawlers" or "spiders," and their page-gathering activity was called "crawling" or "spidering" the Web.

Crawler-based search engines proliferated in 1994. Many of the early search engines were the result of academic or corporate research projects. Two popular engines were the World Wide Web Worm, created by Oliver McBryan at the University of Colorado, and WWW JumpStation, by Jonathon Fletcher at the University of Stirling in the U.K. Neither lasted long: Idealab purchased WWWWorm and transformed it into the first version of the GoTo search engine. JumpStation simply faded out of favor as two other search services launched in 1994 gained popularity: Lycos and Yahoo!.

Michael Mauldin and his team at the Center for Machine Translation at Carnegie Mellon University created Lycos (named for the wolf spider, Lycosidae lycosa, which catches its prey by pursuit, rather than in a web). Lycos quickly gained acclaim and prominence in the Web community, for the sheer number of pages it included in its index (1.5 million documents by January 1995) and the quality of its search results. Lycos also pioneered the use of automated abstracts of documents in search results, something not offered by WWW Worm or JumpStation.

Also in 1994, two graduate students at Stanford University created "Jerry's Guide to the Internet," built with the help of search spiders, but consisting of editorially selected links compiled by hand into a hierarchically organized directory. In a whimsical acknowledgment of this structure, Jerry Wang and David Filo renamed their service "Yet Another Hierarchical Officious Oracle," commonly known today as Yahoo!.

Table 1.1 A Timeline of Internet Search Technologies

Year	Search Service
1945	Vannevar Bush Proposes "MEMEX"
1965	Hypertext Coined by Ted Nelson
1972	Dialog—First Commercial Proprietary System
1986	OWL Guide Hypermedia Browser
1990	Archie for FTP Search, Tim Berners-Lee creates the Web
1991	Gopher: WAIS Distributed Search
1993	ALIWEB (Archie Linking), WWWWander, JumpStation, WWWWorm
1994	EINet Galaxy, WebCrawler, Lycos, Yahoo!
1995	Infoseek, SavvySearch, AltaVista, MetCrawler, Excite
1996	HotBot, LookSmart
1997	NorthernLight
1998	Google, InvisibleWeb.com
1999	FAST
2000+	Hundreds of search tools

In 1995 Infoseek, AltaVista, and Excite made their debuts, each offering different capabilities for the searcher. Metasearch engines—programs that searched several search engines simultaneously—also made an appearance this year (see Chapter 3 for more information about metasearch engines). SavvySearch, created by Daniel Dreilinger at Colorado State University, was the first metasearch engine, and MetaCrawler, from the University of Washington, soon followed.

From this point on, search engines began appearing almost every day. As useful and necessary as they were for finding documents, Web search engines all shared a common weakness: They were designed for one specific task—to find and index Web documents, and to point users to the most relevant documents in response to keyword queries. During the Web's early years, when most of its content consisted of simple HTML pages, search engines performed their tasks admirably. But the Internet continued to evolve, with information being made available in many formats other than simple text documents. For a wide variety of reasons, Web search services began to fall behind in keeping up with both the growth of the Web and in their ability to recognize and index non-text information—what we refer to as the Invisible Web.

To become an expert searcher, you need to have a thorough understanding of the tools at your disposal and, even more importantly, when to use them. Now that you have a sense of the history of the Web and the design philosophy that led to its universal adoption, let's take a closer look at contemporary search services, focusing on their strengths but also illuminating their weaknesses.