

---

# Introduction

If you're like most people, you have a love-hate relationship with search engines and Web directories. You love them, because the Web has become an integral part of daily life, and these pathfinders are crucial guides that help you navigate through an exploding universe of constantly changing information. Yet you also hate them, because all too often they fail miserably at answering even the most basic questions or satisfying the simplest queries. They waste your time, they exasperate and frustrate, even provoking an extreme reaction, known as "Web rage," in some people. It's fair to ask, "What's the problem here? Why is it so difficult to find the information I'm looking for?"

The problem is that vast expanses of the Web are completely invisible to general-purpose search engines like AltaVista, HotBot, and Google. Even worse, this "Invisible Web" is in all likelihood growing significantly faster than the visible Web that you're familiar with. It's not that the search engines and Web directories are "stupid" or even badly engineered. Rather, they simply can't "see" millions of high-quality resources that are available exclusively on the Invisible Web.

So what is this Invisible Web and why aren't search engines doing anything about making it visible? Good question.

There is no dictionary definition for the Invisible Web. Several studies have attempted to map the entire Web, including parts of what we call the Invisible Web. To our knowledge, however, this book represents

the first comprehensive effort to define and map the Invisible Web. We have consulted with numerous Web search experts and developers of major search engines, and have found little consensus among the professional Web search community regarding the cartography of the Invisible Web. Nonetheless, during the course of our research for this book, a relatively clear picture of the properties and boundaries of the Invisible Web has gradually emerged. The picture is constantly shifting in the currents of new and improved technology, but nonetheless paints a portrait that we feel serves as an accurate snapshot of the Invisible Web today.

## The Invisible Web in a Nutshell

The first challenge for the Web searcher is to understand that the Invisible Web exists in the first place. Your interest in this book puts you well on the way. If your searching experience has been limited to the general-purpose Web search tools like Yahoo! and Google, you will soon see that you have been accessing only a small fraction of “Web accessible” information. Many people—even those “in the know” about Web searching—make many assumptions about the scope and thoroughness of the coverage by Web search engines that are simply untrue.

In a nutshell, the Invisible Web consists of material that general-purpose search engines either cannot or, perhaps more importantly, *will not* include in their collections of Web pages (called *indexes* or *indices*). The Invisible Web contains vast amounts of authoritative and current information that’s accessible to you, using your Web browser or add-on utility software—but you have to know where to find it ahead of time, since you simply cannot locate it using a search engine like HotBot or Lycos.

Why? There are several reasons. One is technical—search engine technology is actually quite limited in its capabilities, despite its tremendous usefulness in helping searchers locate text documents on the Web. Another reason relates to the costs involved in operating a comprehensive search engine. It’s expensive for search engines to locate Web resources and maintain up-to-date indices. Search engines must also cope with unethical Web page authors who seek to subvert their indexes with millions of bogus “spam” pages—pages that, like their unsavory e-mail kin, are either junk or offer deceptive or misleading information. Most of the major engines have developed strict

guidelines for dealing with spam, which sometimes has the unfortunate effect of excluding legitimate content.

These are just a few of the reasons the Invisible Web exists. This book takes a detailed look at the nature and extent of the Invisible Web, and offers pathfinders for accessing the valuable information it contains. The bottom line for the searcher is that understanding the Invisible Web and knowing how to access its treasures can save both time and frustration, often yielding high-quality results that aren't easily found any other way.

## What to Expect from This Book

To truly understand what the Invisible Web is, and why it exists, it's important to have a clear understanding of the visible Web and how general-purpose search engines work.

We've designed this book to fit the needs of both novice and advanced Web searchers. If you're new to Web searching, Part I provides essential background information on the design and structure of the Internet, its history and evolution, and the various tools available to help information seekers find what they're looking for. Throughout Part I, we gradually reveal the Invisible Web by describing the structure and operation of the visible Web, and by illustrating the limitations of Web search tools and their gaps in coverage.

If you're a relatively skilled searcher who's already familiar with the nuances of the Web, you can cut to the chase and start with Chapter 3, which begins a detailed exploration of the Invisible Web. Part II, beginning with Chapter 9, is an annotated guide to the best of the Invisible Web. We've selected resources for this section from a broad range of categories that illustrate the high quality of information available on the Invisible Web.

In Chapter 1, *The Internet and the Visible Web*, we trace the development of the Internet and many of the early tools used to locate and share information via the Net. We show how the limitations of these relatively primitive tools ultimately spurred the popular acceptance of the Web. As Tim Berners-Lee, creator of the Web, has written, "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." This historical background, while fascinating in its

own right, lays the foundation for understanding why the Invisible Web could arise in the first place.

Chapter 2, *Information Seeking on the Visible Web*, offers a detailed look at the two predominant Web search services: search engines and Web directories. We examine their strengths and weaknesses, and show how, even though they are useful for finding information on the visible Web, they cannot fully access the riches of the Invisible Web. This chapter discusses the challenges faced by the builders of search engines and directories, and the compromises and tradeoffs they must make that have a direct bearing on what's ultimately included—and excluded—from your search results.

Several prominent studies have determined that search engines simply perform an inadequate job of finding and indexing Web pages. While it is true that search engines do not have comprehensive coverage of the Web, the material they miss is not necessarily part of the Invisible Web. In Chapter 3, *Specialized and Hybrid Search Tools*, we discuss alternative search tools that can help the searcher locate information that, while not part of the Invisible Web, is still difficult if not impossible to find using general-purpose search engines and directories. These specialized and hybrid search tools include targeted directories and crawlers, metasearch engines, value-added search services, “alternative” search tools, and fee-based Web services. We describe and provide examples of all of these types of tools, omitting traditional proprietary database services, which are beyond the scope of the book.

The paradox of the Invisible Web is that it's easy to understand why it exists, but it's very hard to actually define or describe in concrete, specific terms. Nonetheless, that's exactly what we attempt to do in Chapter 4, *The Invisible Web*. In this chapter, we define the Invisible Web, and delve into the reasons why search engines can't “see” its content. We also discuss the four different “types” of invisibility, ranging from the “opaque” Web, which is relatively easy to access, to the truly invisible Web, which requires both determination and specialized finding aids to access its treasures.

In Chapter 5, *Visible or Invisible?*, we get down to the brass-tacks of how to recognize Invisible Web content on your own. We'll show you how to identify Invisible Web pages by looking for telltale signs that signal problems for search engines. We'll also show you how to differentiate between Invisible Web resources and specialized search engines and directories by using a number of comparative case studies.

Although the focus of this book is on the valuable resources found on the Invisible Web, we are not advocating that you abandon the general-purpose search tools you now use. Quite the opposite! In Chapter 6, *Using the Invisible Web*, we discuss why and when to use the Invisible Web to make your Web searching time more efficient by selecting the best available search tool for each particular task. In many respects, searching with a general-purpose search engine is like using a shotgun, whereas searching with an Invisible Web resource is more akin to a taking a highly precise rifle-shot approach. It's only by thinking carefully about your quarry that you'll be able to select your appropriate search "weapon."

Though there are many technical reasons why major search engines don't index the Invisible Web, there are also "social" reasons having to do with the validity, authority, and quality of online information. Because the Web is open to everybody and anybody, a good deal of its content is published by non-experts—or even worse, by people with a strong bias that they seek to conceal from readers. As mentioned earlier, search engines must also cope with millions of bogus "spam" pages. No matter whether you're searching the visible or Invisible Web, it's important to always maintain a critical view of the information you're accessing. Chapter 6 covers some important techniques for assessing the validity and quality of online information. We also present some excellent resources for keeping current with the rapid growth of the Invisible Web.

Chapter 7, *Case Studies*, presents eight scenarios that demonstrate both the power of Invisible Web resources, and why general-purpose search tools simply fail miserably at finding the materials used in the examples. In each case study, we attempt not only to show how search tools function, but also to illustrate the problem-solving approach the searcher uses to satisfy an information need.

The Invisible Web's value and rapid growth have attracted the attention of some skilled researchers who are working to make it more accessible by general search tools. In Chapter 8, *The Future: Revealing the Invisible Web*, we take a brief look at some of the more interesting approaches and projects likely to illuminate portions of the Invisible Web in coming years.

The directory section of the book begins with Chapter 9, *The Best of the Invisible Web*. This chapter describes a number of exceptional pathfinder sites that provide links to high-quality Invisible Web content. The remaining chapters make up a directory of more than 1,000

Invisible Web sites hand-selected by the authors. Each chapter focuses on a specific topic or subject to help you quickly pinpoint the resources you need for a wide range of information needs. The directory includes resources that are informative, of high quality, and contain worthy information from reliable information providers that are not visible to general-purpose search engines. We give precedence to resources that are freely available to anyone with Web access.

As an added bonus, we have made this directory available online at the companion Web site for this book, [www.invisible-web.net](http://www.invisible-web.net). The online directory includes the most up-to-date annotations and links for each resource, and is continually updated to include new Invisible Web resources as we locate them.

Throughout the book, we include sidebars debunking commonly held beliefs about search engines and searching the Web that are simply untrue. These “Web Search Myths” can lead to poor or even disastrous results for the unwary searcher. They also can lead to false assumptions about what is—and is not—part of the Invisible Web.

Although the Invisible Web is a relatively complex subject, our style is informal, seeking to demystify our topic rather than impress the reader with our erudition. By necessity, there are a fair number of technical terms used in the book. Whenever we introduce a technical term that’s particularly important, we also provide an accompanying definition box nearby. The glossary contains complete definitions of all of these important terms, as well as all other technical terms used in the book.

## Boldly Go Where No Search Engine Has Gone Before

By now, you’re probably convinced that the Invisible Web is an incredibly valuable resource for serious searchers. It is, but there are a number of things you should keep in mind as you set out to explore the Web’s hidden reaches.

The Invisible Web is huge, vaguely defined, and incorporates databases with a wide variety of interfaces. This means that there is a fairly significant learning curve for getting comfortable with what’s available. Don’t despair if it seems overwhelming at first! Learning to use

Invisible Web resources is just like learning any new and valuable resource. Though it may seem like second nature now, when you first learned how to look a word up in the dictionary or find a number in the telephone book, it took time, patience, and practice, too.

And remember—there are numerous exceptions to the rules. In this book we have done our best to generalize in a way that does not make incorrect assumptions. We understand and acknowledge that there are inconsistencies in some aspects of the book. We don't view these as "gotchas"—rather, we feel exceptions to the rules illustrate the richness inherent in the Web as a whole.

The Invisible Web holds incredibly valuable resources for the searcher. Journeys into the Invisible Web lead not only to treasures that aren't easily located, but often provide the pleasure and satisfaction experienced by early explorers who led expeditions into regions of the world marked *Terra Incognita* on early maps. In *The Invisible Web*, our goal is to provide you with a detailed map of a vast expanse of cyberspace that is still relatively uncharted territory, allowing you to boldly go where no search engine has gone before.