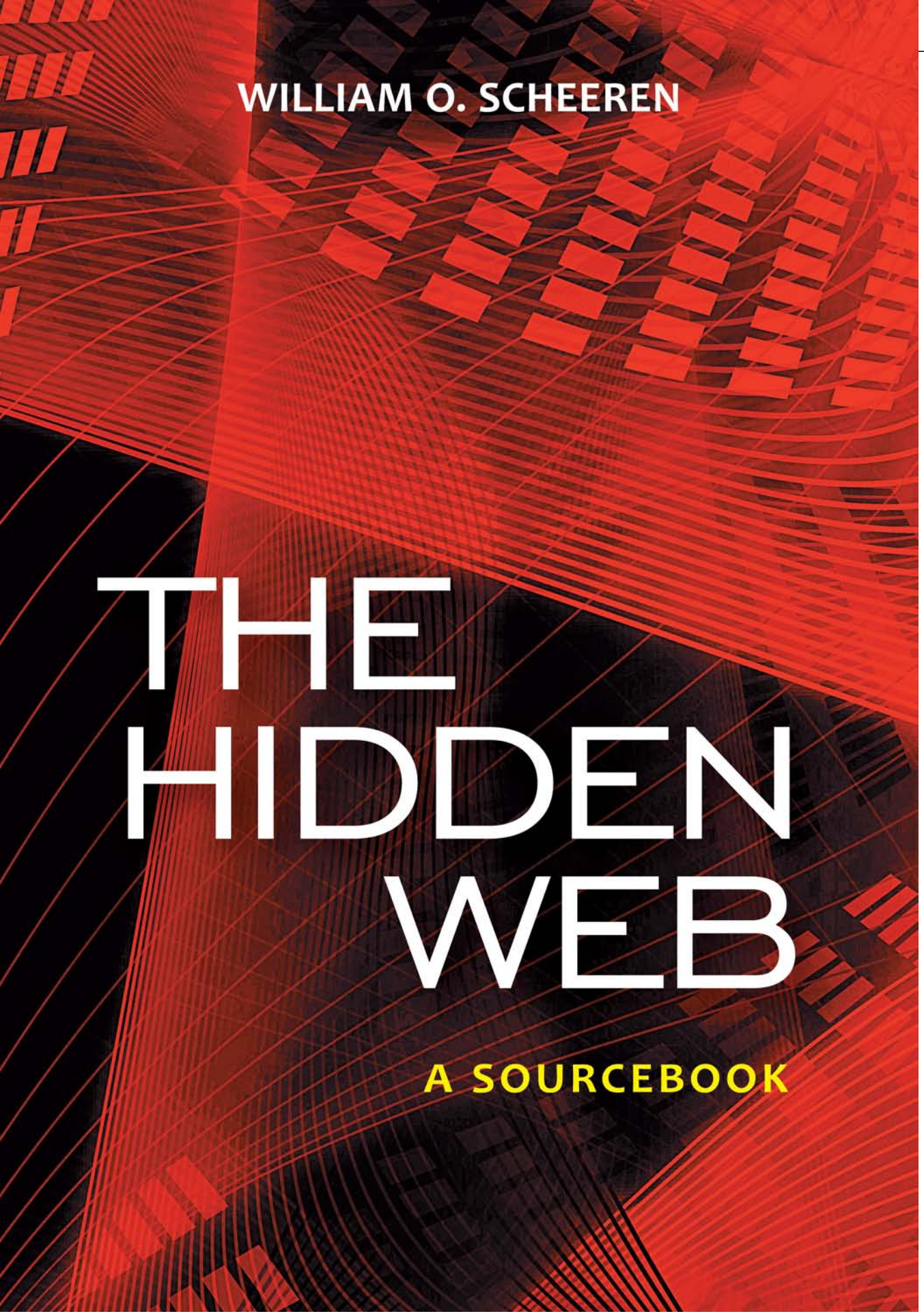
WILLIAM O. SCHEEREN
THE HIDDEN WEB
A SOURCEBOOK

THE HIDDEN WEB

THE HIDDEN WEB

A Sourcebook

William O. Scheeren

LIBRARIES UNLIMITED

AN IMPRINT OF ABC-CLIO, LLC
Santa Barbara, California • Denver, Colorado • Oxford, England

Library of Congress Cataloging-in-Publication Data

Scheeren, William O.
The hidden Web : a sourcebook / William O. Scheeren.
pages cm
Includes bibliographical references and index.
ISBN 978–1–59884–627–0 (pbk.) — ISBN 978–1–59884–628–7 (ebook) 1. Invisible Web. 2. Internet addresses—Directories. 3. Internet searching. 4. Database searching. 5. Internet in school libraries. I. Title.
ZA4237.S34 2012
025.042′2—dc23 2011042977

ISBN: 978–1–59884–627–0
EISBN: 978–1–59884–628–7

16 15 14 13 12 1 2 3 4 5

This book is also available on the World Wide Web as an eBook.
Visit www.abc-clio.com for details.

Libraries Unlimited
An Imprint of ABC-CLIO, LLC

ABC-CLIO, LLC
130 Cremona Drive, P.O. Box 1911
Santa Barbara, California 93116-1911

This book is printed on acid-free paper

Manufactured in the United States of America

Contents

1

What Is the Invisible Web and Why Is It Important to Librarians?

Throughout my years as a school librarian, the biggest challenge faced by both my students and I was finding the best information that most specifically answered their research questions. Prior to the proliferation of the Internet and other electronic resources, it was easier to locate information in the school library because it was either in the library or it was not. If the information was in the library, the students and I could easily locate it; if it was not in the library, there were relatively few alternatives as to how to obtain the information. Some libraries were members of a consortium or had cooperative agreements for interlibrary loan; some of these welcomed school libraries into their membership. If you were fortunate enough to be in an area that had college or large public libraries, students were generally welcome to use the facilities, albeit with many restrictions such as not being allowed to borrow materials. The reality was that often these opportunities were not available for reasons of time and distance.

The move to electronic resources in school libraries was slower than for public and academic libraries especially when they served a larger clientele and had larger budgets. The use of electronic resources in schools can be traced on a continuum beginning with the use of commercial databases such as LexusNexis, Orbit, and Dialog. Using these databases was a real breakthrough in information gathering, but they were just the initial steps towards where we are now. First, for many school districts these databases were prohibitively expensive. They were subscription services, but in addition to the subscription price, they also charged by the search. Second, both because of the cost and the complexity of the search process, the searching had to be done by the librarian. No school district was going to pay the fees involved with the databases and allow students to search or browse indiscriminately. Certainly, the use of these

subscription databases was a step forward in the information gathering process. However, they still remained very expensive for individual schools to purchase and if the school district was small, buying a license for a database remained very expensive.

One of the solutions for the use of databases in every school came about when states began to use state funds to purchase databases for use in every library within the state. However, this type of resource was limited to the states whose legislators would fund such purchases. Training was also something that needed to be shared with librarians within the state so that they could teach their patrons how to access the databases. This opportunity became a starting point for many students who could access these databases while they were in school or if they went to their public libraries, even if they did not have access to a computer at home.

For national access to information, it was the introduction and the proliferation of the Internet and the World Wide Web (WWW) that changed information-gathering behavior forever in school libraries as well as in other types of libraries. It is impossible to overstate how important the Internet has been to libraries, to librarians, and to students. Who among us does not remember our early experiences with the Internet: Dial-up modems, terminal-like screen displays, and really limited amounts of data? There was even a modicum of humor as we did Gopher searches through the main "Gopher Hole" at the University of Minnesota, whose athletic teams are known as the "Golden Gophers." The reality was, however, that we, as school librarians, had a very powerful way to access and deliver information to our clients. Unfortunately, we did not have a clue or even a suspicion how to use it effectively at the time.

Let us move forward to the school library of the twenty-first century. Technology has completely changed what we do. The computer has become more and more powerful and less and less expensive. Students and librarians are adept at finding exactly the information they need by searching the Internet. Or are they as adept as we think (and hope!)? Ask students and teachers how they search for and locate information on the Internet and the overwhelming answer is that they use a search engine. Furthermore, many users are such basic searchers that the search engines often return 10,000,000 or more results, many of which turn out to be irrelevant to their information needs.

Even the most sophisticated searcher, no matter what search engine they use, is only touching the surface of the information available on the Internet because they are not able to search the Invisible Web. The Invisible Web is also known as the Deep Web, Deepnet, or the Hidden Web. The Invisible Web contains information that cannot be accessed by general purpose search engines such as Google (Devine and Egger-Sider, 2009, p. 3). Several authors have compared the use of today's search engines that search for information on the Internet to fishing trawlers who trawl for fish on or near the surface of the ocean. They

get a lot of information but are just touching the surface of what is out there. The terminology or concept of the Invisible Web has been in existence for over 15 years, but it is still little known by students or librarians.

In an informal survey of my students in a master's degree program, I asked how they search for information on the Internet. The vast majority of them said they used broad search terms in a general purpose search engine. A few stated they used search limiters when using the general search engines and also made use of the college's subscription databases. None of my students had even heard of the Invisible Web. When I explained the type of information that could be found on the Internet, they agreed unanimously that they could not understand why they did not know about it. Unfortunately, many librarians are as unaware of the Invisible Web and what it contains as are my students.

Before we proceed with a general discussion of what is contained in the Invisible Web, we need to establish that the Invisible Web is far, far larger than the Visible Web. I could provide some numbers about the size of the Internet and about how many Web sites exist, but that information would be out-of-date as soon as I put the numbers on the page. What we do know is that there are billions of Web sites in existence and that for every Web site that can be accessed using a standard search engine such as Google, there are five to ten that exist on the Invisible Web and cannot be located using basic search engines. Furthermore, the Invisible Web is growing at a far greater rate than is the Visible Web (Devine and Egger-Sider, 2009, p. xxi).

That now brings us around to a general discussion and description of what types of information are contained in the Invisible Web. All users should keep in mind, however, that while some items may be invisible using certain search engines, they may be part of the Visible Web because other search engines can locate them (Sherman and Price, 2007, p. 84). In general, though, the following types of items are only accessible through the Invisible Web (Devine and Egger-Sider, 2009, pp. 3 and 9).

1. Database Content. One would suppose that this would be part of the Visible Web, but search engines have almost no success locating database content because they are dynamically generated by querying the database. It is impossible to replicate that type of search either by using a general search engine or by entering the URL for the search results page in a browser because results were dynamically generated and then, the search form was reset. In other words, a search for information such as comparisons of literacy rates among countries is created by dynamically generated HTML pages that cannot be recreated without using the original database.
2. Deep Web Sites. This category sounds as though we are talking in riddles, but sites known as Deep Web sites generally are so data intensive that search engines such as Yahoo and Google cannot find their information.

3. As a general rule, some file formats cannot be located using standard search engines. This category includes any non-HTML formats. This will be discussed in more depth in Chapter 3, What Is in the Invisible Web, Can It Be Searched, and Why Use the Invisible Web?
4. Forms. This is closely related to the database results discussed above. Forms normally reset once the form is completed and submitted and therefore, even using the URL of the form in your browser before it is submitted will be unsuccessful. This again has to do with the dynamic content of forms.
5. Very Current Material. The most current information is often not available through standard search engines because the vast amount of this current information is often beyond the ability of most standard search engines to assimilate and index it.
6. Ephemeral Information. In contrast to information that remains on the Internet for what seems to be forever is ephemeral information. This is information that is either time-sensitive or so short-lived that standard search engines never have it available.
7. Gray Literature. This is a category of non-technical government information that is rarely found using standard search engines and resides on this enormous Invisible Web.

The last issue addressed in this chapter is why the Invisible Web should matter to librarians. In reality, the answer to this is far simpler than finding information on the Invisible Web: it allows librarians and students to find the best, most precise information and data that suits their information needs. It is up to the librarian to learn what is available on the Invisible Web and be prepared to share this knowledge with faculty so that they can encourage its use with their students. It is also important that librarians help students use the Invisible Web when they have a reference question or a research assignment which requires the information found there.

2

How Search Engines Work (or Do Not Work) with the Invisible Web

In Chapter 1, the concept that material in the Invisible Web cannot generally be accessed or retrieved using standard search engines such a Google or Yahoo was discussed. Keep in mind, however, that not all search engines return the same things as others even when the exact same search is done. Furthermore, meta search engines such as Dogpile use multiple search engines to search the Internet and return extremely comprehensive hit lists. What all of this means is that if you search "Shakespeare" in Yahoo, Google, and Dogpile, each search engine will have many of the same results, but there will be differences in the results lists for each of the three search engines.

It is perhaps important to note that general or standard search engines will not find material in the Invisible Web, and it cannot be found for two reasons. First, as powerful as many search engines are, they have technology limitations that prevent them from getting to the Invisible Web. Second, it is extremely expensive to develop and maintain comprehensive, general purpose search engines. Those search engines look at literally billions of Web sites and attempt to organize them into some manageable whole. It is just not financially practical, therefore, for a comprehensive general purpose search engine to delve into the Invisible Web even if it could (Sherman and Price, 2007, p. xiii).

Let's now reexamine the types of files that are typically not found by standard search engines and are therefore a part of the Invisible Web. In general, search engines do not "play well" with material that is not text-based. Web pages that are primarily video, audio, or images—in other words non-text-based material—are rarely accessible through the standard search engine. There

are some specific files formats within these general file types that search engines cannot handle. These are:

- PDF or Postscript formats unless they come from Google.
- Flash. Of course, Apple and iPads also have issues with this file format.
- Shockwave.
- Programs. Actually all executable files.
- Compressed files such as .zip files.

The difficulty with indexing these types of files is that they are not HTML text and standard search engines generally do not choose to index them, mostly for financial reasons (Sherman and Price, 2007, p. 58).

In the larger picture, these types of files make up a small percentage of the material found in the Invisible Web. The much larger amount of material found on the Invisible Web is one of two types: (1) single Web pages or (2) database information. Single Web pages are generally Web pages created by individual users. The information contained on these Web pages is sometimes valuable, but they cannot be located by standard search engines because there are no links for the Web crawlers that are at the heart of search engines to locate the page.

The second large category of material located on the Invisible Web is database information which can generally be further divided into three categories. The first is database material that is designed for the needs of individual users. This data is often generated by forms and is contained in relational databases. Standard Web search engines cannot fill out the required information in interactive forms and therefore, even if you have an exact URL of the search, it will not return the data. The second type of data found in databases in the Invisible Web is streaming or real-time content. Because there is so much of it and it changes so rapidly, standard Web search engines just cannot keep up with this content. The third type is dynamically generated content. This is similar to the first item discussed relative to database information on the Invisible Web (Sherman and Price, 2007, p. 60-61).

Throughout this chapter we have considered why search engines cannot *access* material that is contained in the Invisible Web, but we have not discussed the technical aspects of how search engines *find* material on the Web. Most people do not need to know much more than how to enter a search term in a search engine; the search engine searches the Web and then produces a results list of Web sites that the search engine has determined meet the search criteria. This is a gross oversimplification, however. If it were that easy, then simple searches would always be sufficient. Let's now consider some problems (shortcomings) noted with general purpose search engines. These problems can be categorized into three general areas: (1) retrieval limits, (2) search strategy problems, and (3) evaluation issues.

RETRIEVAL LIMITS

1. The Web simply has too much material for any one search engine to find and index everything. Tough decisions have to be made both from a practical and a financial standpoint. It goes without saying that sometimes the information you want is not indexed and is therefore not available on the Visible Web.
2. Much of the information returned in a result list is irrelevant to the searcher. Why does this occur? It occurs generally because of poor or inexact search strings. For example, if a searcher is interested in material about William Shakespeare and enters the search term "Shakespeare," not only will the results list have information about William Shakespeare, but also about Shakespeare fishing equipment, Shakespeare festivals, etc.
3. Again, there are the search engine limitations that have been discussed earlier in this chapter. It is not technically or financially feasible to design or create a search engine that finds everything on every topic every time.
4. People do not realize that everything on the Web is not free. When the author was working as a school librarian and the Web was in its infancy, a member of our board of school directors stated that there would soon be no need for brick-and-mortar school libraries because everything was on the Web and it was all free. Sadly, this has not proved to be true.

SEARCH STRATEGY PROBLEMS

1. Poor search skills are closely related to item two above: irrelevant data or result lists. If searchers are not trained in search skills, often the result list will not be relevant. Training users in search skills is a key part of information literacy instruction for school librarians.
2. Nearly all search engine users consider themselves to be advanced or experienced searchers. The reality is that few are experienced or diligent enough searchers to take advantage of such advanced search engine features as Boolean Operators or search limiters.
3. Some searchers always use the same search engine. While this develops a certain amount of skill in searching, it ignores the fact that different search engines almost always return different result lists. Routinely using a meta search engine can alleviate this issue to some degree.
4. Again related to several items that have been discussed previously, many searchers do the simplest kind of searches. In our previous example of Shakespeare, this is an example of a simple search. A more sophisticated searcher might try to limit the search terms and use Boolean search limits.
5. Psychological studies have shown that Web users have a distinct aversion to scrolling through multiple screen result lists. When a result list returns many (perhaps even thousands) of result screens, it is almost inevitable that some useful results will be missed.

6. As a corollary of item five, there is often a high level of frustration when viewing numerous links to material. It often becomes overwhelming to the searcher.
7. Some of the value of items found on a result list lessens when there are thousands or millions of results.
8. Searchers are impatient people as witness the studies that show they are not willing to scroll through several pages of a results list. After a certain point, they just feel they are wasting time.
9. Searchers often do not use the Help feature of the search engine. This is generally the easiest, most accessible assistance and most searchers ignore it.
10. Quality sites are returned in searchers' result lists. Often however, searchers are more impressed with the volume of information on a site rather than the quality of the information.

EVALUATION ISSUES

1. General search engines do not evaluate the information that is returned in results lists. It is incumbent of the searcher to determine what is the best information.
2. Many, in fact most commercial Web sites contain advertising. After all, this is how many Web sites are paid for. Unfortunately, general search engines see this advertising and will include them in result lists. This is how information about Shakespeare fishing equipment becomes intermixed with information about William Shakespeare.
3. Tons of inappropriate sites are out there on the Web. Web search engines often pick them up (Devine and Egger-Sider, 2009, p. 29–31).

The final topic that will be discussed in this chapter is a series of what might be called Web searching myths. These are items that most people believe to be true but are, in fact, false.

1. Everything worth finding is already on the Web. If it can't be found on the Web, it is not worth finding. This is obviously not true. If it were true there would be need to update Web sites or to create new Web sites.
2. Google and other general search engines search the entire Web. We already know this is not correct or there would be no such thing as the Invisible Web. The Invisible Web contains at least 500 times as much material as the Visible Web.
3. The best information is in the first ten results. As mentioned above search engines do not rate results in any way. Therefore, it follows that there can be no hard and fast rule that the best material is in the first ten items of a result list.

4. Searching is easy. This can be true, but good searching is hard. Anyone can type in a search term but it takes skill to construct a search string that returns what the searcher needs.
5. Everything important on the Web is free. How many ways can we say no to this? In fact, the best material is often in subscription databases.
6. Everything on the Web is truthful, authentic, and accurate. Suffice to say that anyone can post anything to the Web at any time. (Devine and Egger-Sider, 2009, p. 4).

3

What Is in the Invisible Web, Can the Invisible Web Be Searched, and Why Use the Invisible Web?

Chapters 1 and 2 examined what the Invisible Web is, why the Invisible Web is a valuable resource for the school librarian, and how search engines do or do not work with the Invisible Web. In Chapter 3, the final chapter before we move to the chapters that deal specifically with Invisible Web sites of value to school librarians, we will take a more in-depth look at (1) what is in the Invisible Web, (2) determining if the Invisible Web can be searched, and, the biggest question of all, (3) why we should use the Invisible Web. In addition, we will examine what are considered to be the most useful Invisible Web categories. Finally, we will enumerate what cannot be found on the Web, using either the Visible or the Invisible Web.

The second question mentioned above is the one that will be answered first: can the Invisible Web be searched? At the gross level, the answer is "no." Basic search engines such as Google and Yahoo can locate Invisible Web tools, but they are not able to extract information from these tools (Devine and Egger-Sider, 2009, p. 147). The basic search engine is able to get to an Invisible Web tool such as a database containing census information, but you are not able to mine this information directly from the database using the basic search engine. You must go to the database on the Invisible Web to mine for and extract specific information.

The Visible Web is also known as the Surface Web. If we go back to the analogy of accessing the Visible Web as like trawling the ocean for fish at a shallow depth, then the term "Surface Web" becomes obvious. If we go back also to our discussion about what is in the Invisible Web—single Web pages, databases, current information, and forms—we can then compare the contents of the Surface Web with those of the Invisible Web. Bear in mind that the figures