The “Deep Web”

Deep Web – Those billions of webpages that are technically accessible via the Internet, but which are not indexed by search engines and therefore cannot be located just by means of a search engine search.

Why are Some Web Pages "Invisible" to Search Engines?

Important Point: Most “websites” are not “invisible” to search engines, but for sites referred to as “Deep Web” sites”, the majority of pages within the site may be “invisible”, i.e., not findable by search engine searches.

1. Web pages that search engines could find and include, but choose not to.

   Example: The page is too deep in the sites and search engine producers don't expend their resources to go that deeply. ("depth of crawl")

   Particularly true of extremely large sites and of sites that are primarily "free home page sites." Though this can cause many good pages to be missed, it may also enhance the average usefulness of sites included in search engines.

Even these sites “can” appear in web search engines if another site links to them or if someone “submits” the site to a search engine.

2. Web pages that a search engine could include, but can't find.

   "Disconnected pages."

   The way search engines proactively find pages is by going to pages they already know about and looking for links to pages they don't already know about ("crawling").

   If there are no links to a page, the search engine can't find it by crawling. On the right is an example of a page that might not be found because no one linked to it.
3. Pages that a search engine has been asked not to index
Pages where, for privacy reasons, the creator of the sites "asks" search engine crawler programs ("robots") not to index the page. This is done by using "robot exclusion protocols".

4. Pages that search engines can't access because the search engine doesn't have the necessary password.

5. Pages that a search engine cannot index because it can't "interpret" them because of the format.

Why can't they?

The entire page is in a file format that their crawlers can't handle

Images - A search engine can't tell a tree from a tractor unless there is associated text telling it so. (software is coming along to do so, but is not yet ready for search engines, or vice-versa.)

Sounds - Likewise
6. Pages that a search engine cannot index because the information is stored in databases and the search engine cannot get beyond the search box on the page.

This is the REALLY BIG category.

Why can’t search engines access these kinds of pages?

They don’t know what to put in the search box in order to get beyond the main page of the site.

This also includes pages that are created "on the fly" from relational databases.

Just because these kinds of material are "invisible" to search engines does not mean you can’t get to them.

What You Can Do to Find Deep Web Pages

Know the tools that identify Deep Web sites:

- Resource Guides
- Searching for terms such as “database”, e.g.,

  real estate assessment virginia database
Tools for Identifying Deep Web Sources of Interest

Resource Guides

These sites are the best starting place for finding the most important sites (both invisible and visible) for a particular subject area.

“Resource guides” are collections of links on a specific topic. If there are any significant database sites for that topic, those sites will usually show up somewhere on the resource guide.

Trick to easily find resource guides:

Do a search on the topic of interest and the word “resources”, e.g.,

- education resources
- "lesson plan" resources
- history resources

Though sites will almost never say "This Website is an Invisible Web site", it doesn't matter. If the resource guides point to a site that is of interest to you, once you go to the site, you will see how to get to the site's content.

One very good collection of Deep Web examples is found at Complete Planet (completeplanet.com)

Ran Hock 2008

(This document may be copied freely for classroom use. RH)