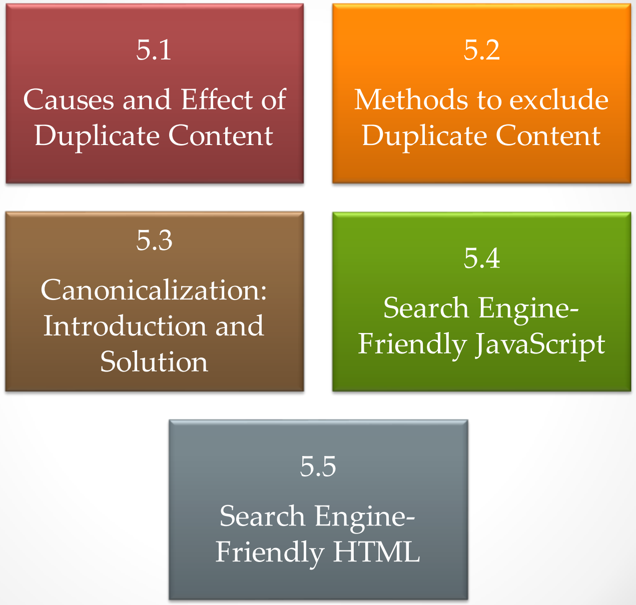
Unit - 5

Duplicate Content, SE-Friendly HTML and JavaScript



5.1 Causes and Effect of Duplicate Content

5.2 Methods to exclude Duplicate Content

5.3 Canonicalization: Introduction and Solution

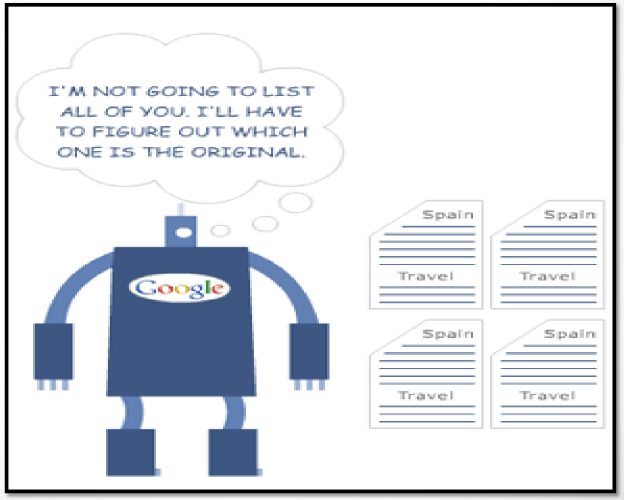
5.4 Search Engine-Friendly JavaScript

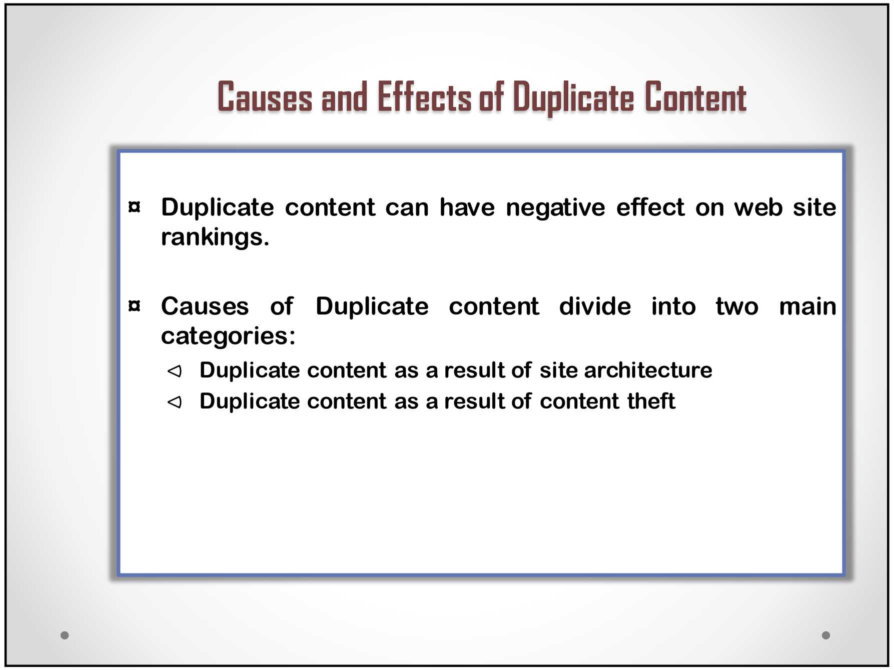
5.5 Search Engine-Friendly HTML

Duplicate Content

- ¤ It is a web content that is either exactly duplicated or substantially similar to content located at different URLs.
- ¤ Duplicate content clearly does not contain anything original.
- ¤ When there are multiple pieces of identical content on the Internet, it is difficult for search engines to decide which version is more relevant to a given search query.
- ¤ Search engines employ sophisticated algorithms that detect such content and filter it out from results.
- ¤ Indexing and processing duplicate content wastes the storage and computation time of search engine.

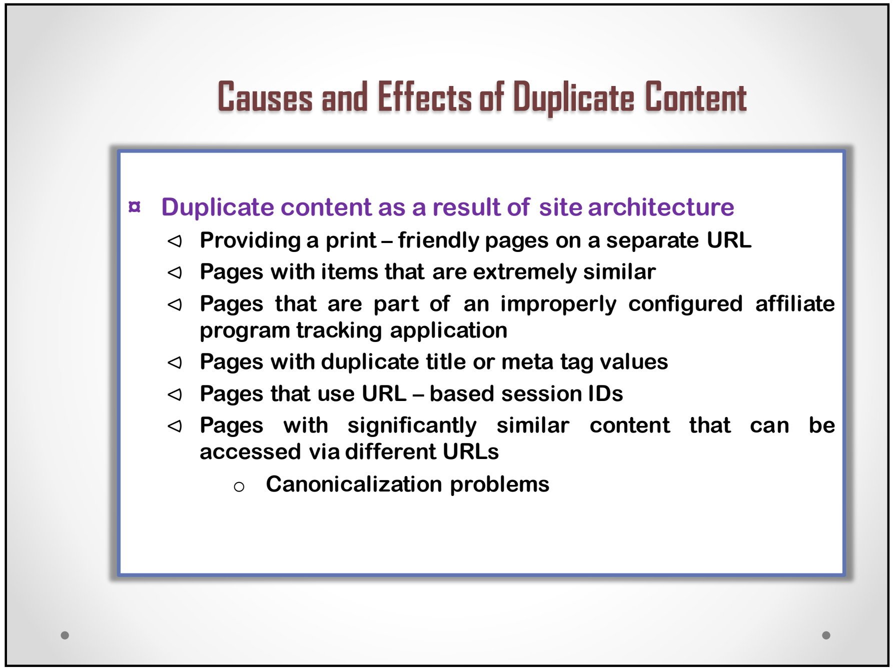
Duplicate Content


I'M NOT GOING TO LIST ALL OF YOU. I'LL HAVE TO FIGURE OUT WHICH ONE IS THE ORIGINAL.
Google
Spain
Travel
Spain
Travel
Spain
Travel
Spain
Travel



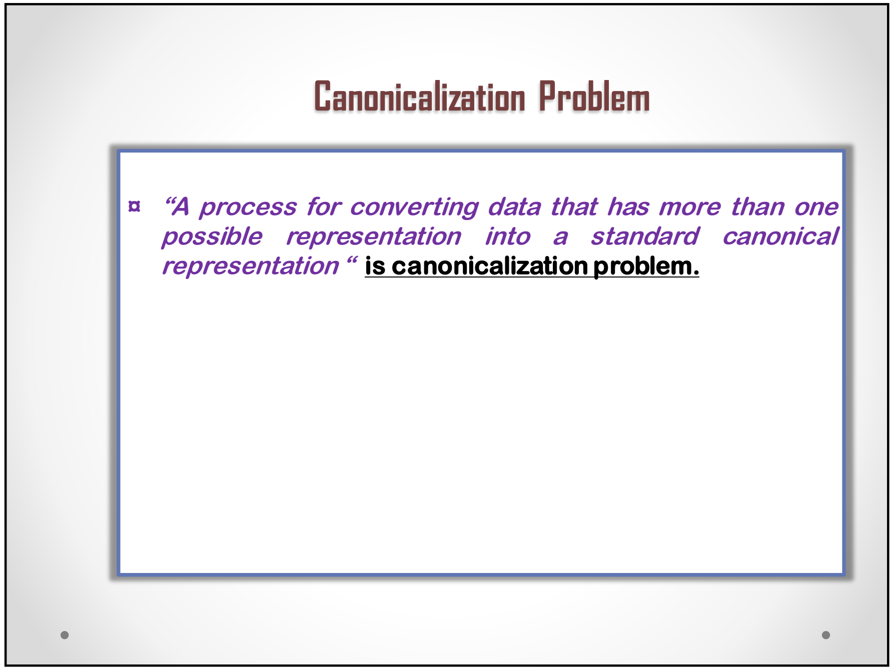
Causes and Effects of Duplicate Content

- Duplicate content can have negative effect on web site rankings.
- Causes of Duplicate content divide into two main categories:
  - Duplicate content as a result of site architecture
  - Duplicate content as a result of content theft



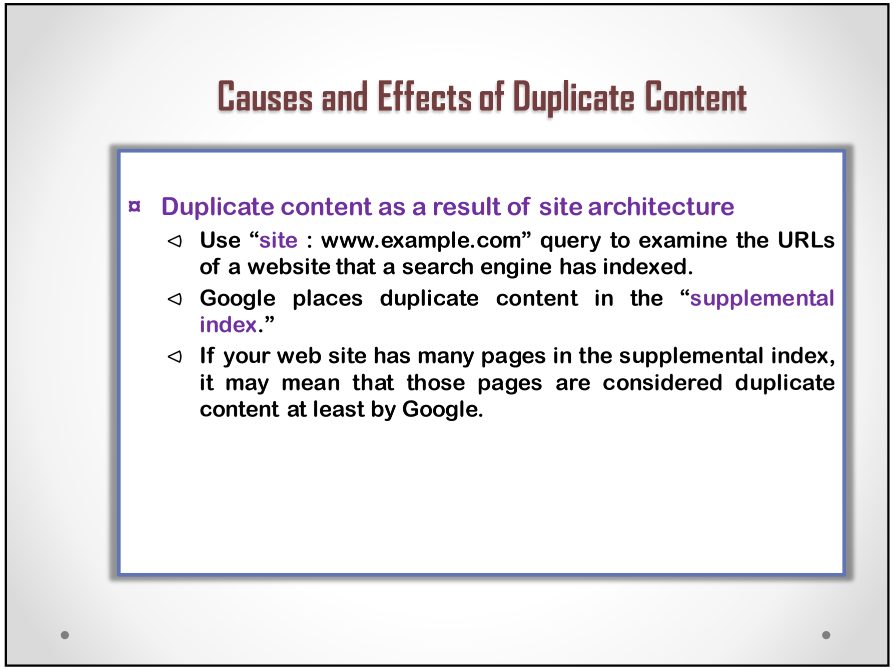
Causes and Effects of Duplicate Content

- Duplicate content as a result of site architecture
  - Providing a print – friendly pages on a separate URL
  - Pages with items that are extremely similar
  - Pages that are part of an improperly configured affiliate program tracking application
  - Pages with duplicate title or meta tag values
  - Pages that use URL – based session IDs
  - Pages with significantly similar content that can be accessed via different URLs
    - Canonicalization problems



Canonicalization Problem

- *"A process for converting data that has more than one possible representation into a standard canonical representation "* **<u>is canonicalization problem.</u>**



Causes and Effects of Duplicate Content

- Duplicate content as a result of site architecture
  - Use "site : www.example.com" query to examine the URLs of a website that a search engine has indexed.
  - Google places duplicate content in the "supplemental index."
  - If your web site has many pages in the supplemental index, it may mean that those pages are considered duplicate content at least by Google.

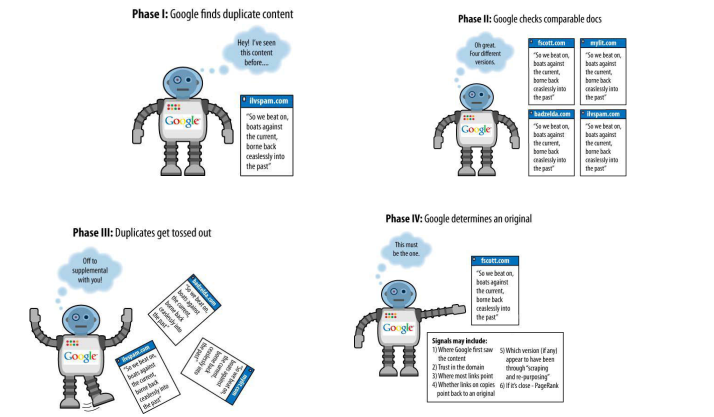
Causes and Effects of Duplicate Content

- Duplicate content as a result of content theft
  - It creates a similar problem for search engines, which attempt to filter duplicate content from search result.
  - It will sometime make the wrong assumption as to which instance of the content is the original, authoritative one.
  - CopyScape is a service that helps to find content thieves by scanning for similar content contained by other pages.
  - Major search engines have procedure to alert of stolen content.

Causes and Effects of Duplicate Content

- Duplicate content as a result of content theft
  - URLs with the directions for the major search engines:
    - ✓ Google: http://www.google.com/dmca.html
    - ✓ Yahoo!: http://docs.yahoo.com/info/copyright/copyright.html
    - ✓ MSN: http://search.mas.com/docs/siteowner.aspx?t=SEARCH_WEBMASTER_CONC_AboutDMCA.htm

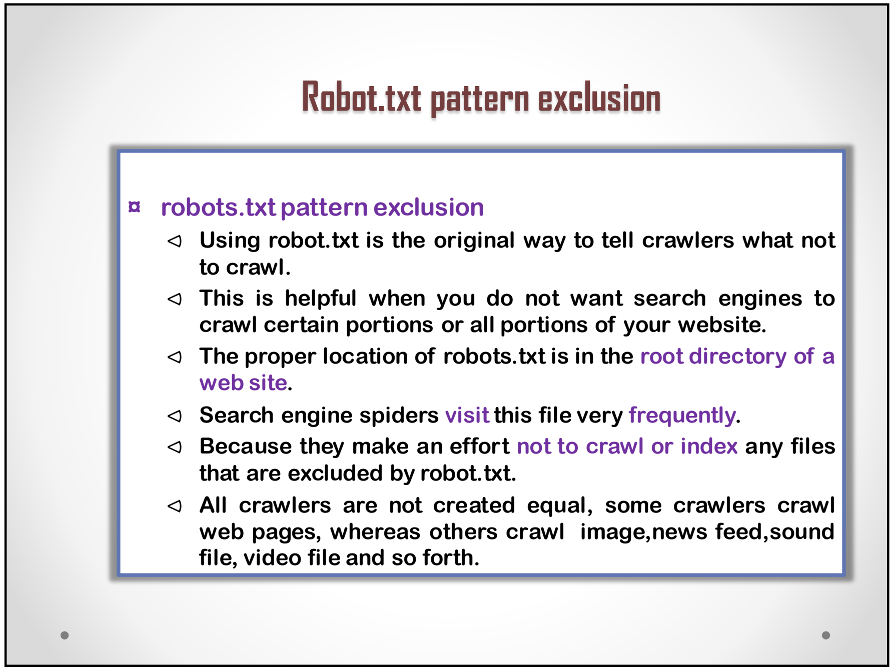
How Search Engine Identify Duplicate Content


Phase I: Google finds duplicate content
Hey! I've seen this content before....
ilvspam.com
"So we beat on, boats against the current, borne back ceaslessly into the past"
Google
Phase II: Google checks comparable docs
Oh great. Four different versions.
fscott.com
mylit.com
badzelda.com
ilvspam.com
Phase III: Duplicates get tossed out
Off to supplemental with you!
Phase IV: Google determines an original
This must be the one.
fscott.com

Signals may include:
1) Where Google first saw the content
2) Trust in the domain
3) Where most links point
4) Whether links on copies point back to an original
5) Which version (if any) appear to have been through "scraping and re-purposing"
6) If it's close - PageRank

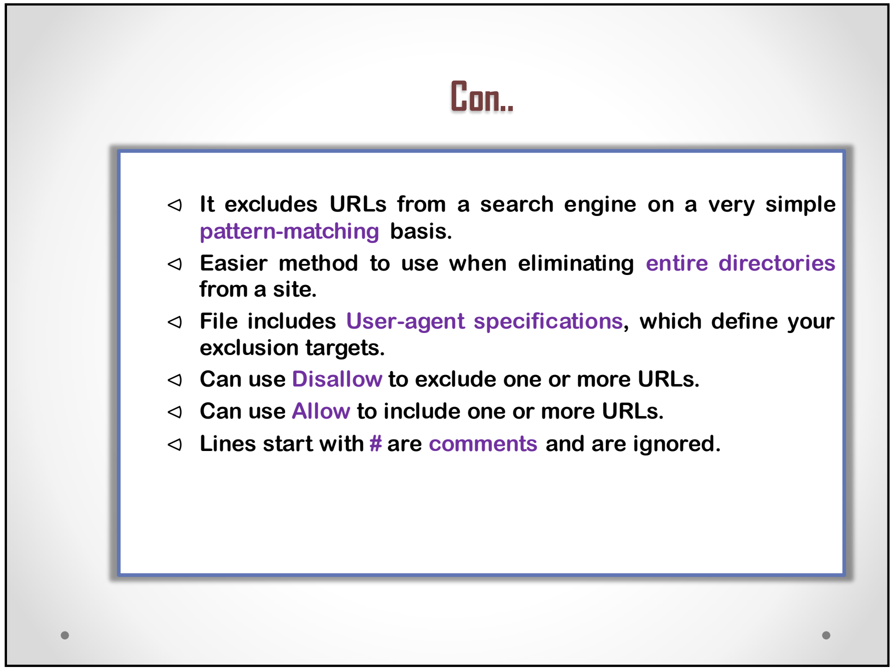
Excluding Duplicate Content

- In case of duplicate content on your site, you can remove it by altering the architecture of a web site.
- But sometime duplicate content included because of business rules that drive the web site.
- To address this, you can exclude it from the view of a search engine.
  - Use canonical tag.
  - Use robots.txt pattern.
  - Use robots meta tag.



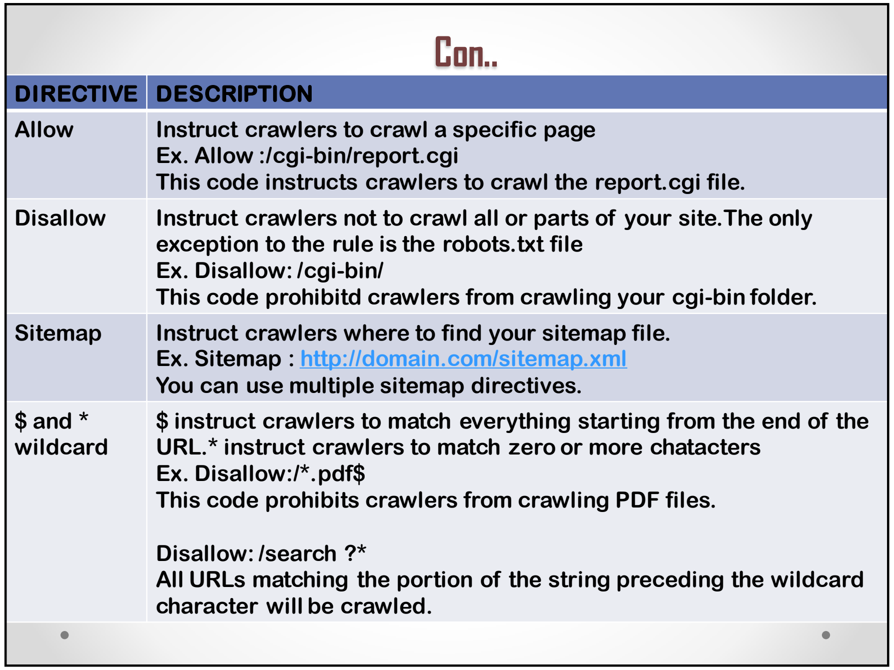
Robot.txt pattern exclusion

- robots.txt pattern exclusion
  - Using robot.txt is the original way to tell crawlers what not to crawl.
  - This is helpful when you do not want search engines to crawl certain portions or all portions of your website.
  - The proper location of robots.txt is in the root directory of a web site.
  - Search engine spiders visit this file very frequently.
  - Because they make an effort not to crawl or index any files that are excluded by robot.txt.
  - All crawlers are not created equal, some crawlers crawl web pages, whereas others crawl image,news feed,sound file, video file and so forth.



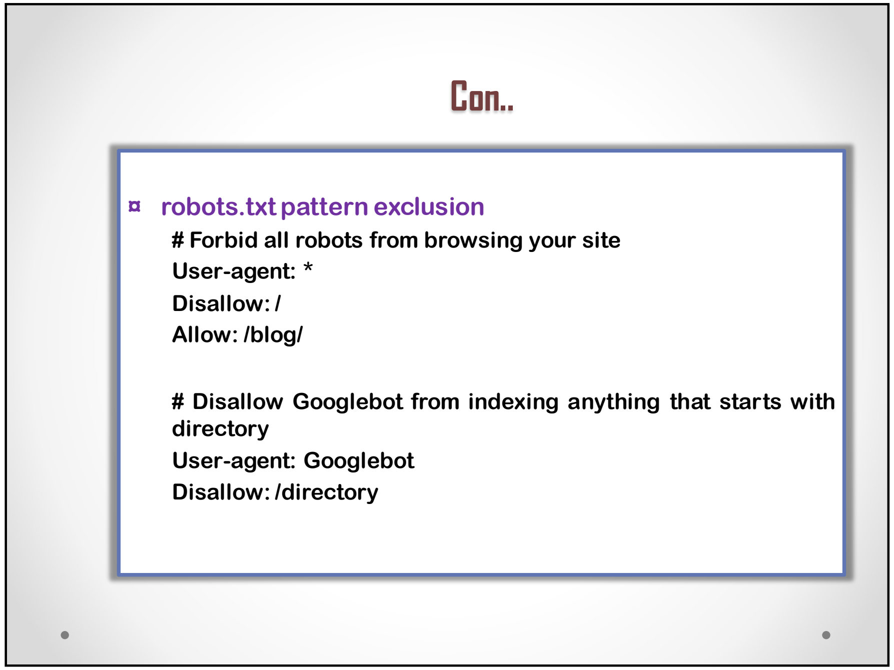
Con..

- It excludes URLs from a search engine on a very simple pattern-matching basis.
- Easier method to use when eliminating entire directories from a site.
- File includes User-agent specifications, which define your exclusion targets.
- Can use Disallow to exclude one or more URLs.
- Can use Allow to include one or more URLs.
- Lines start with # are comments and are ignored.



Con..

DIRECTIVE	DESCRIPTION
Allow	Instruct crawlers to crawl a specific page Ex. Allow :/cgi-bin/report.cgi This code instructs crawlers to crawl the report.cgi file.
Disallow	Instruct crawlers not to crawl all or parts of your site.The only exception to the rule is the robots.txt file Ex. Disallow: /cgi-bin/ This code prohibitd crawlers from crawling your cgi-bin folder.
Sitemap	Instruct crawlers where to find your sitemap file. Ex. Sitemap : http://domain.com/sitemap.xml You can use multiple sitemap directives.
$ and * wildcard	$ instruct crawlers to match everything starting from the end of the URL.* instruct crawlers to match zero or more chatacters Ex. Disallow:/*.pdf$ This code prohibits crawlers from crawling PDF files. Disallow: /search ?* All URLs matching the portion of the string preceding the wildcard character will be crawled.



Con..

- robots.txt pattern exclusion

```
# Forbid all robots from browsing your site
User-agent: *
Disallow: /
Allow: /blog/

# Disallow Googlebot from indexing anything that starts with directory
User-agent: Googlebot
Disallow: /directory
```

Con..

¤ robots.txt pattern exclusion

```
# Block all robots from tmp and logs directories
User-agent: *
Disallow: /tmp/
Disallow: /logs        # for files called logs

# Block access to all subdirectories that begin with private
User-agent: Googlebot
Disallow: /private*/
```

Robots meta tag

¤ Using the robots meta tag

- ◁ Using it you can exclude any HTML based content from a web site on a page-by-page basis.
- ◁ To exclude a page, place following code in the <head> section of the HTML document you want to exclude:

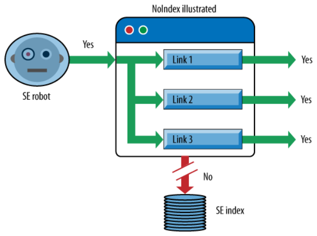
<meta name=“robots” content=“noindex, nofollow” />

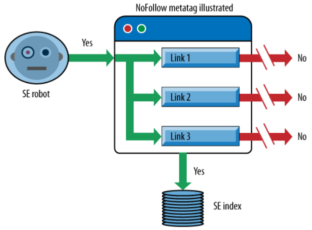
noindex: page should not be indexed

nofollow: none of the link on the page should be crawl

Con..

¤ Using the robots meta tag





Con..

¤ Using the robots meta tag

- ◁ To exclude a specific spider, change “robots” to the name of the spider.
- ◁ To exclude multiple spiders, you can use multiple meta tags.

<meta name=“googlebot” content=“noindex, nofollow” />

Search Engine	User Agent
Google	Googlebot
Yahoo!	Slurp
MSN Search	Msnbot
Ask	Teoma

Con..

- Using the robots meta tag
  - Two technical limitations are associated with this method.
    - ✓ It requires access to the source code of the application.
      - Otherwise it become impossible because the tag must be placed in the webpages generated by the application.
    - ✓ It only works with HTML files, not with clear text, CSS, or binary/image files.

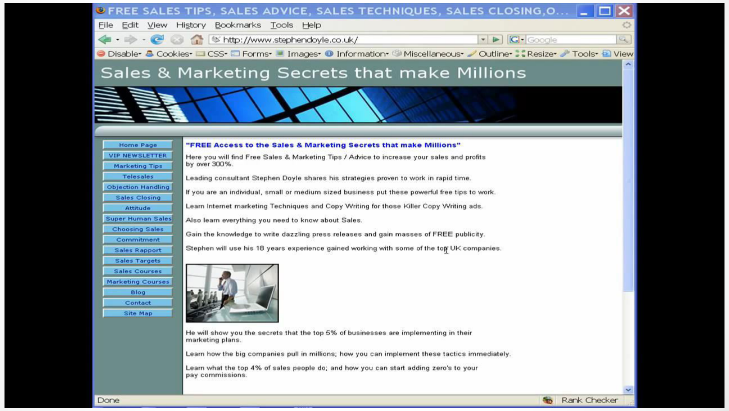
Search Engine – Friendly JavaScript

- Search engine are designed to index content rather than execute application code.
- When it used the wrong way, can degrade a web site's search engine friendliness.
- JavaScript's use in context of the following:
  - DHTML menus
  - Links
  - Popups
  - Crawlable images and graphical text

Search Engine – Friendly JavaScript

- DHTML Menus
  - Many drop-down menus are somewhat spider-friendly, whereas others are not at all.
  - It often present problems for search engines as well.
  - It is wise to provide alternative navigation to all elements listed in the menus.
  - Can do using a set of links at the bottom of the page, a sitemap, or a combination thereof.
  - Because of that also visitors with JavaScript support disabled will be easily able to navigate the site.

Search Engine – Friendly JavaScript



FREE SALES TIPS, SALES ADVICE, SALES TECHNIQUES, SALES CLOSING,O...
File Edit View History Bookmarks Tools Help
http://www.stephendoyle.co.uk/
Google
Disable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools View
Sales & Marketing Secrets that make Millions
Home Page
VIP NEWSLETTER
Marketing Tips
Telesales
Objection Handling
Sales Closing
Attitude
Super Human Sales
Choosing Sales
Commitment
Sales Rapport
Sales Targets
Sales Courses
Marketing Courses
Blog
Contact
Site Map
Done
Rank Checker

"FREE Access to the Sales & Marketing Secrets that make Millions"

Here you will find Free Sales & Marketing Tips / Advice to increase your sales and profits by over 300%.

Leading consultant Stephen Doyle shares his strategies proven to work in rapid time.

If you are an individual, small or medium sized business put these powerful free tips to work.

Learn Internet marketing Techniques and Copy Writing for those Killer Copy Writing ads.

Also learn everything you need to know about Sales.

Gain the knowledge to write dazzling press releases and gain masses of FREE publicity.

Stephen will use his 18 years experience gained working with some of the top UK companies.

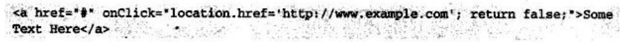
He will show you the secrets that the top 5% of businesses are implementing in their marketing plans.

Learn how the big companies pull in millions; how you can implement these tactics immediately.

Learn what the top 4% of sales people do; and how you can start adding zero's to your pay commissions.

Search Engine – Friendly JavaScript

- JavaScript Links
  - It is any button or text that, when clicked, navigates to another page.
  - Typical JavaScript link looks like:



```
<a href="#" onClick="location.href='http://www.example.com'; return false;">Some
Text Here</a>
```

  - It prevent a search engine spider from following the links, and also prevent users who disable JavaScript from navigating your site.
  - Using them for all navigation may prevent a site from being spidered at all.

Search Engine – Friendly JavaScript

- JavaScript Links
  - If you must use links, provide alternative navigation somewhere else on the site.
  - Same issues would also be specious in navigation involving other client side technologies such as Java applets, AJAX content and Flash.
  - Means any navigation not achieved using a standard anchor (<a>) tag will delay site spidering.

Search Engine – Friendly JavaScript

- Popup Windows
  - Typical method of displaying popups employs JavaScript.
  - Search engine will not spider a page only referred to by JavaScript.
  - Typical popup link looks like this:

```
<a href="#" onClick="window.open('page.html', 'mywindow', 'width=800,
height=600'); return false; " target="_blank">Click here </a>
```

  - You could make the popup spiderable by changing the link to this:

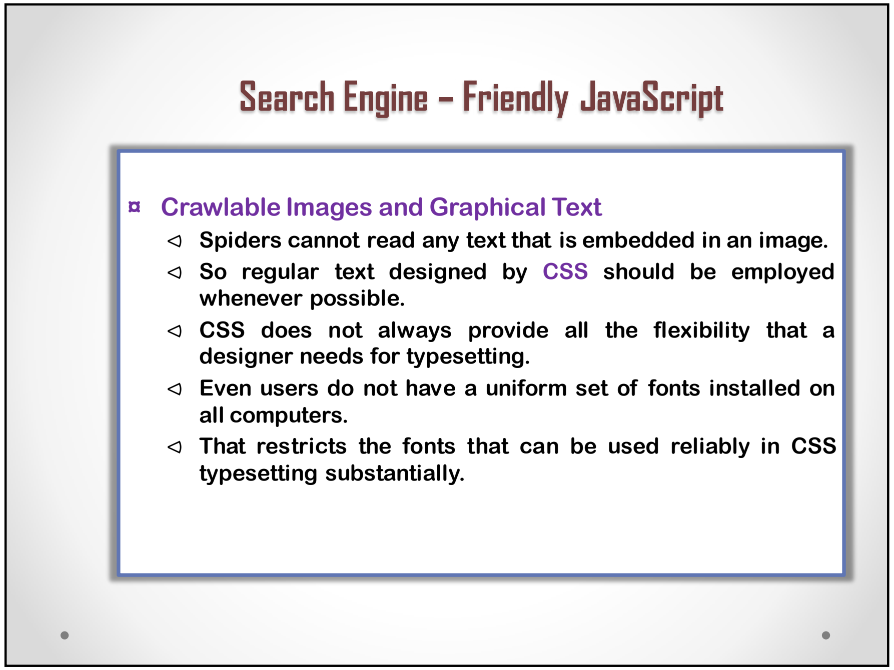
```
<a href="page.html" onClick="window.open('this.href', 'mywindow', 'width=800,
height=600'); return false; " target="_blank">Click here </a>
```

Search Engine – Friendly JavaScript

- DHTML Popup Windows
  - You can place an invisible <div> element at particular location, then use JavaScript events to hide and unhide.

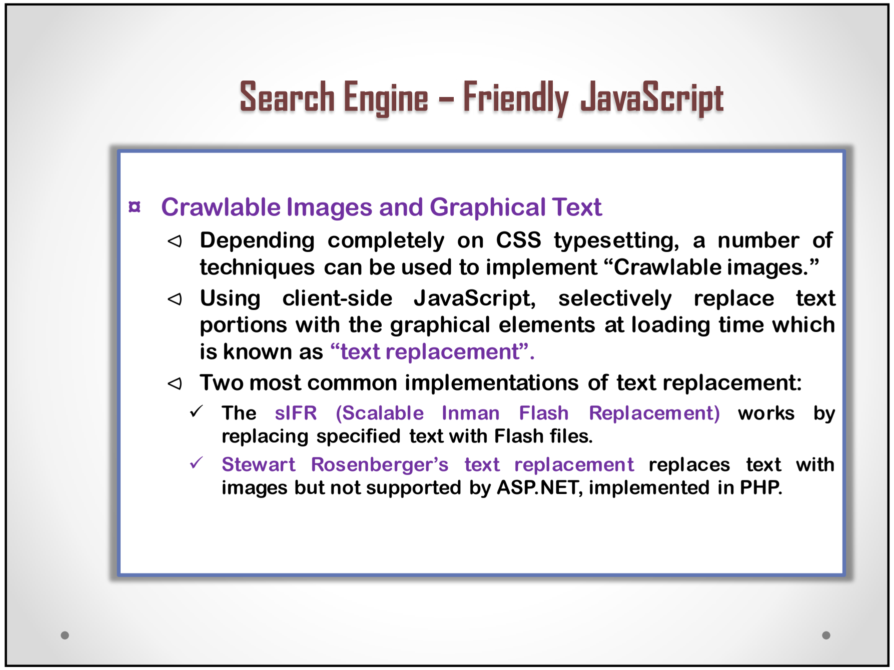
```
<span onmouseover="document.getElementId('dhtml-popup-
test').style.visibility='visible';" onmouseout= "document.getElementId('dhtml-
popup-test').style.visibility='hidden';"> put mouse here </span>

<div style="position:absolute; visibility:hidden; border: 1 px solid black"
id="dtml-popup-test"> this only visible if mouse  is over </div>
```

  - Advantage : Although text is spiderable it may be regarded as invisible on page factor because it is not visible by default.



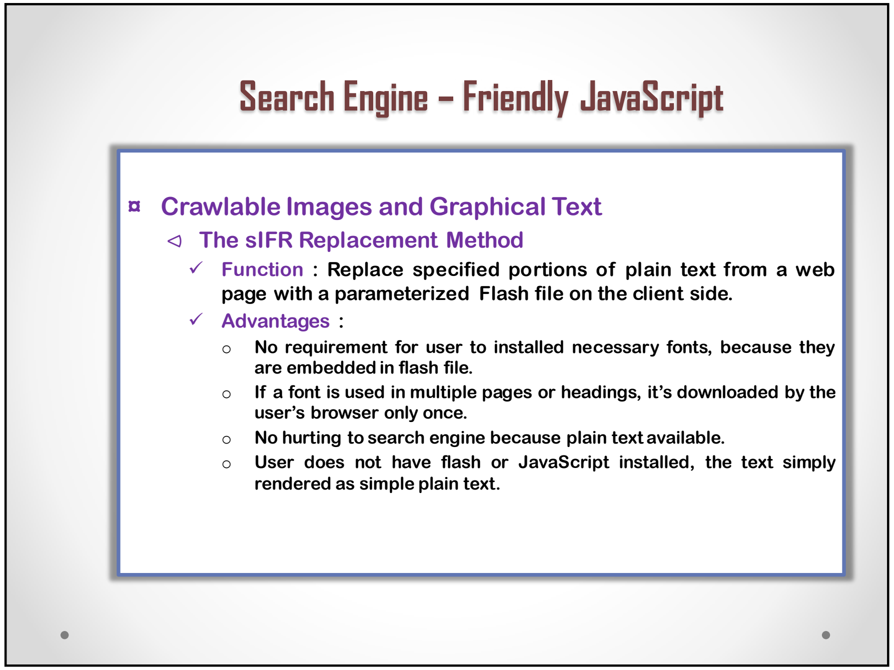
Search Engine – Friendly JavaScript

- Crawlable Images and Graphical Text
  - Spiders cannot read any text that is embedded in an image.
  - So regular text designed by CSS should be employed whenever possible.
  - CSS does not always provide all the flexibility that a designer needs for typesetting.
  - Even users do not have a uniform set of fonts installed on all computers.
  - That restricts the fonts that can be used reliably in CSS typesetting substantially.



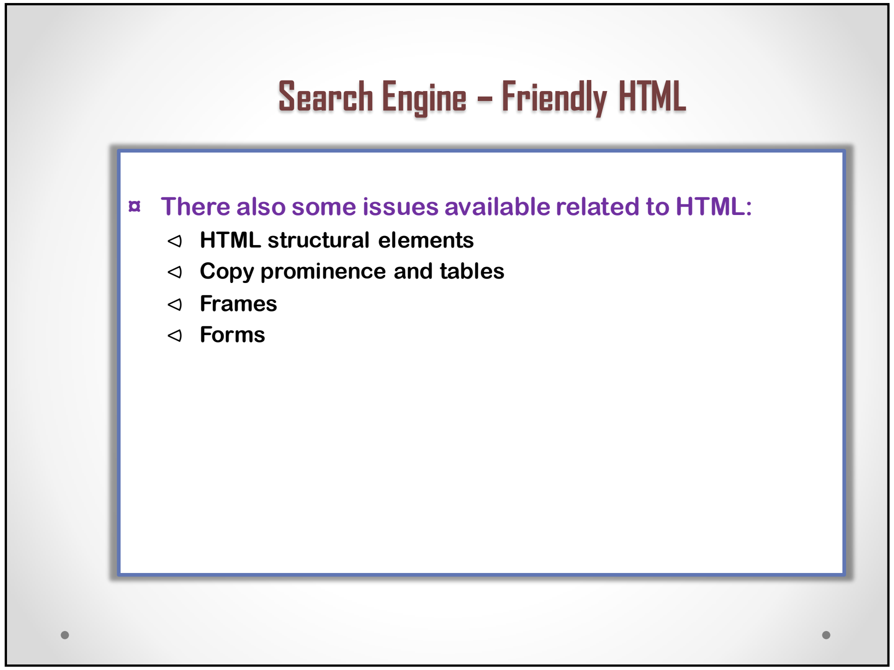
Search Engine – Friendly JavaScript

- Crawlable Images and Graphical Text
  - Depending completely on CSS typesetting, a number of techniques can be used to implement "Crawlable images."
  - Using client-side JavaScript, selectively replace text portions with the graphical elements at loading time which is known as "text replacement".
  - Two most common implementations of text replacement:
    - ✓ The sIFR (Scalable Inman Flash Replacement) works by replacing specified text with Flash files.
    - ✓ Stewart Rosenberger's text replacement replaces text with images but not supported by ASP.NET, implemented in PHP.



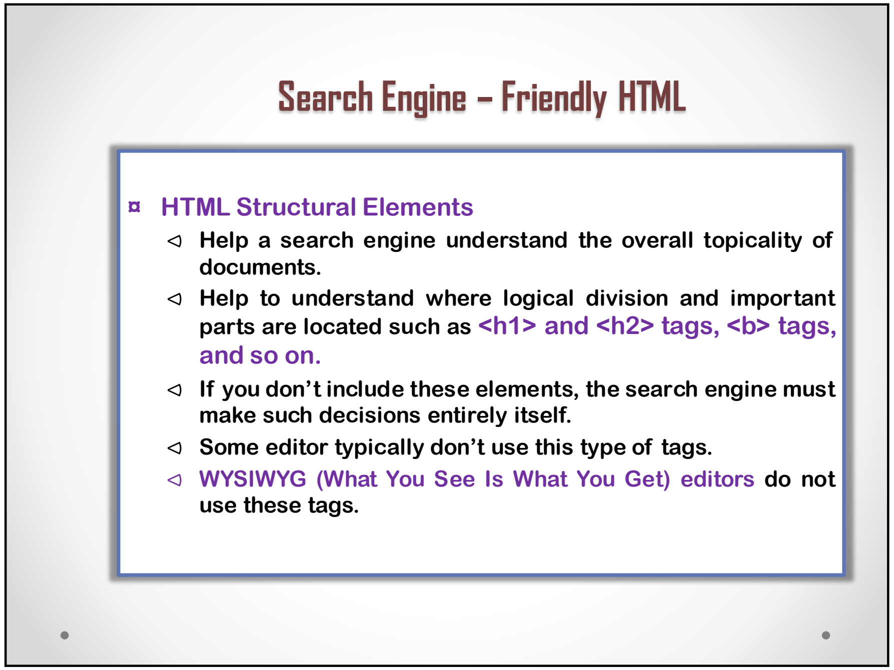
Search Engine – Friendly JavaScript

- Crawlable Images and Graphical Text
  - The sIFR Replacement Method
    - ✓ Function : Replace specified portions of plain text from a web page with a parameterized Flash file on the client side.
    - ✓ Advantages :
      - No requirement for user to installed necessary fonts, because they are embedded in flash file.
      - If a font is used in multiple pages or headings, it's downloaded by the user's browser only once.
      - No hurting to search engine because plain text available.
      - User does not have flash or JavaScript installed, the text simply rendered as simple plain text.



Search Engine – Friendly HTML

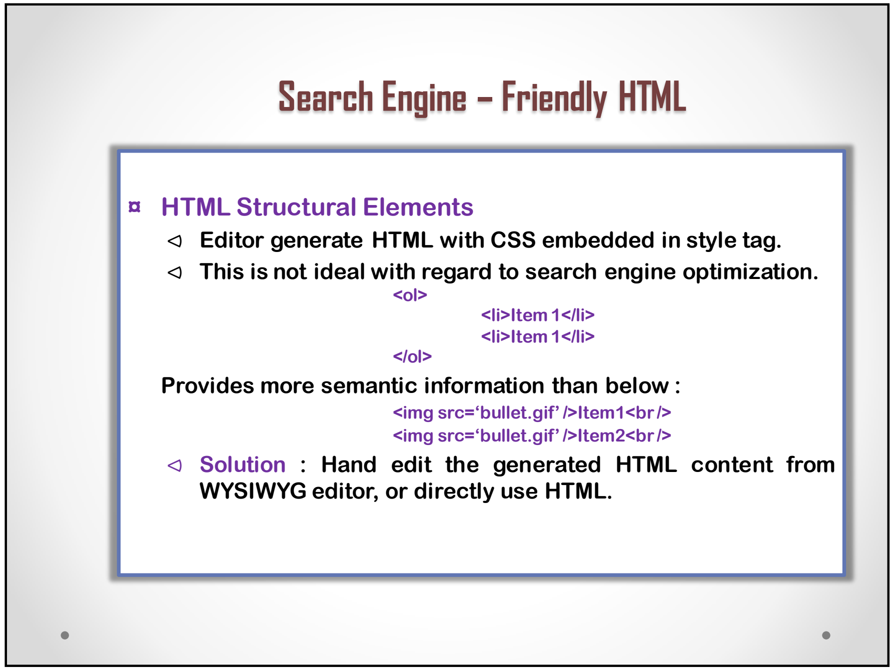
- There also some issues available related to HTML:
  - HTML structural elements
  - Copy prominence and tables
  - Frames
  - Forms



Search Engine – Friendly HTML

¤ HTML Structural Elements

- ◃ Help a search engine understand the overall topicality of documents.
- ◃ Help to understand where logical division and important parts are located such as <h1> and <h2> tags, <b> tags, and so on.
- ◃ If you don't include these elements, the search engine must make such decisions entirely itself.
- ◃ Some editor typically don't use this type of tags.
- ◃ WYSIWYG (What You See Is What You Get) editors do not use these tags.



Search Engine – Friendly HTML

¤ HTML Structural Elements

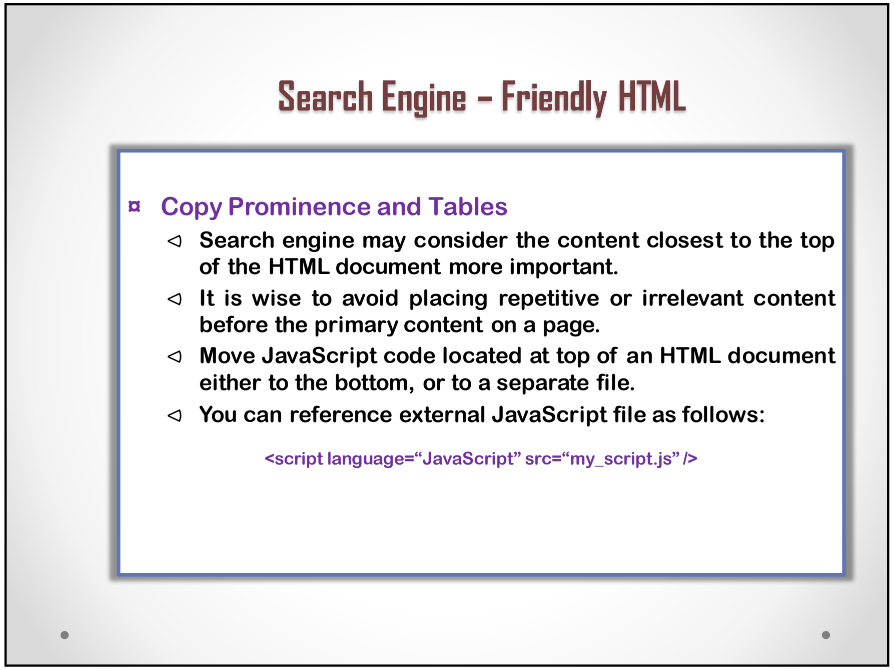
- ◃ Editor generate HTML with CSS embedded in style tag.
- ◃ This is not ideal with regard to search engine optimization.

```
<ol>
        <li>Item 1</li>
        <li>Item 1</li>
</ol>
```

Provides more semantic information than below :

```
<img src='bullet.gif' />Item1<br />
<img src='bullet.gif' />Item2<br />
```

- ◃ Solution : Hand edit the generated HTML content from WYSIWYG editor, or directly use HTML.

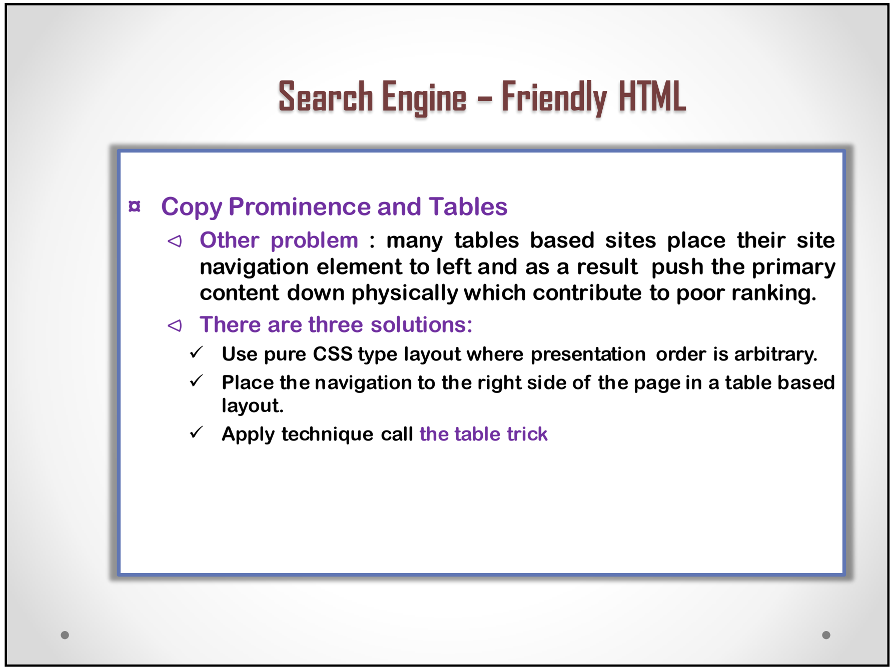


Search Engine – Friendly HTML

¤ Copy Prominence and Tables

- ◃ Search engine may consider the content closest to the top of the HTML document more important.
- ◃ It is wise to avoid placing repetitive or irrelevant content before the primary content on a page.
- ◃ Move JavaScript code located at top of an HTML document either to the bottom, or to a separate file.
- ◃ You can reference external JavaScript file as follows:

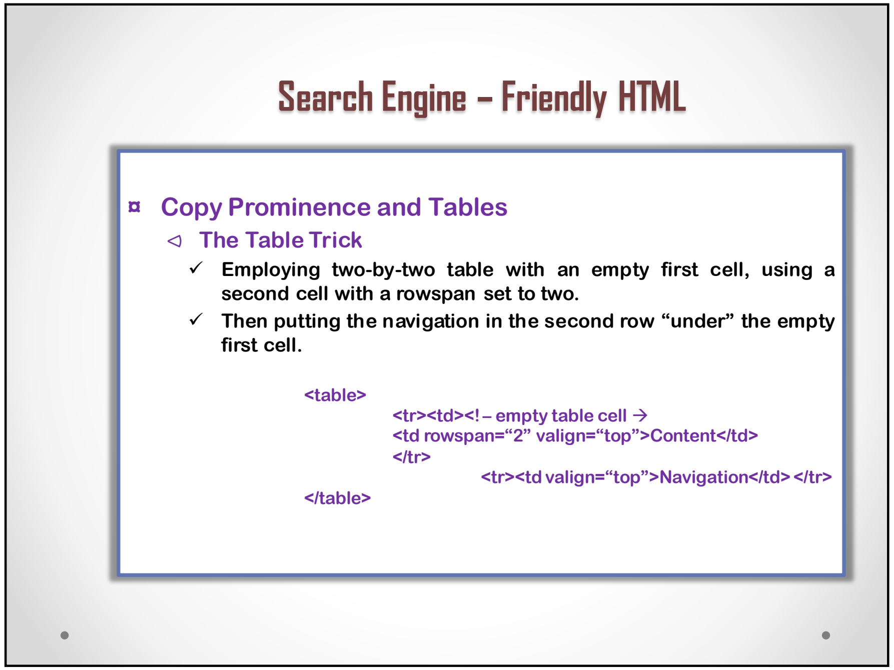
```
<script language="JavaScript" src="my_script.js" />
```



Search Engine – Friendly HTML

¤ Copy Prominence and Tables

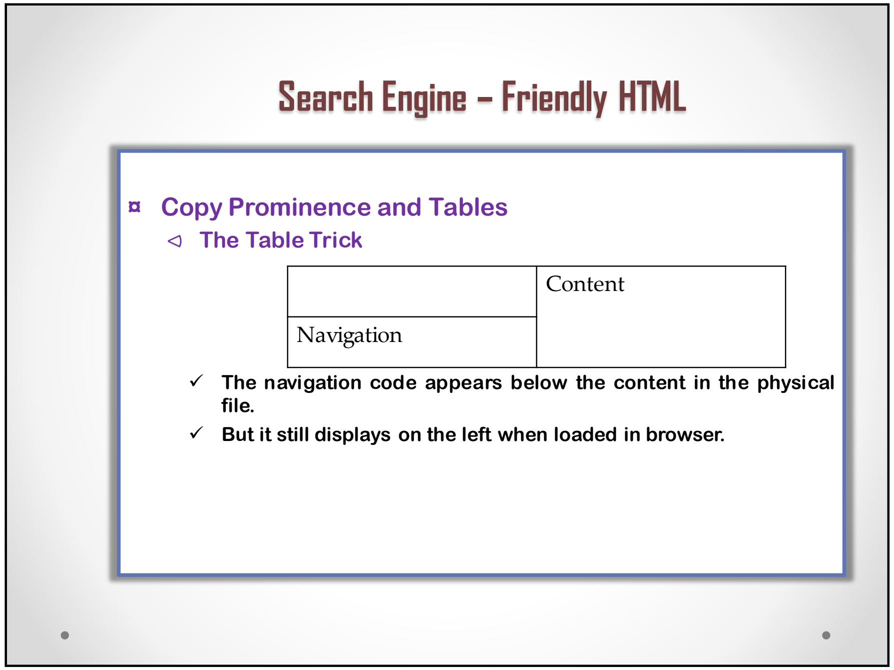
- ◃ Other problem : many tables based sites place their site navigation element to left and as a result push the primary content down physically which contribute to poor ranking.
- ◃ There are three solutions:
  - ✓ Use pure CSS type layout where presentation order is arbitrary.
  - ✓ Place the navigation to the right side of the page in a table based layout.
  - ✓ Apply technique call the table trick



Search Engine – Friendly HTML

- Copy Prominence and Tables
  - The Table Trick
    - ✓ Employing two-by-two table with an empty first cell, using a second cell with a rowspan set to two.
    - ✓ Then putting the navigation in the second row "under" the empty first cell.

```
<table>
            <tr><td><! – empty table cell →
            <td rowspan="2" valign="top">Content</td>
            </tr>
                        <tr><td valign="top">Navigation</td> </tr>
</table>
```

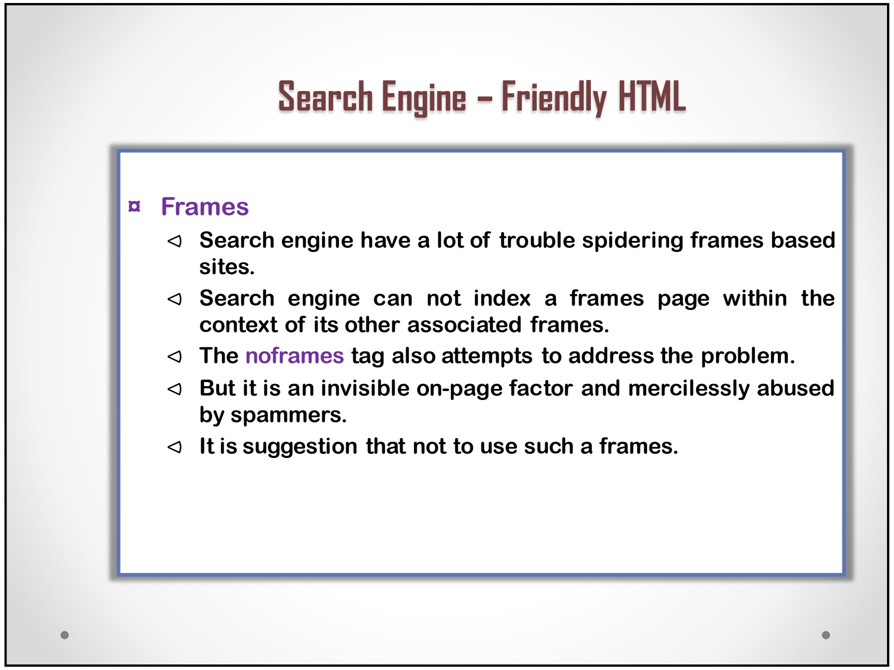


Search Engine – Friendly HTML

- Copy Prominence and Tables
  - The Table Trick

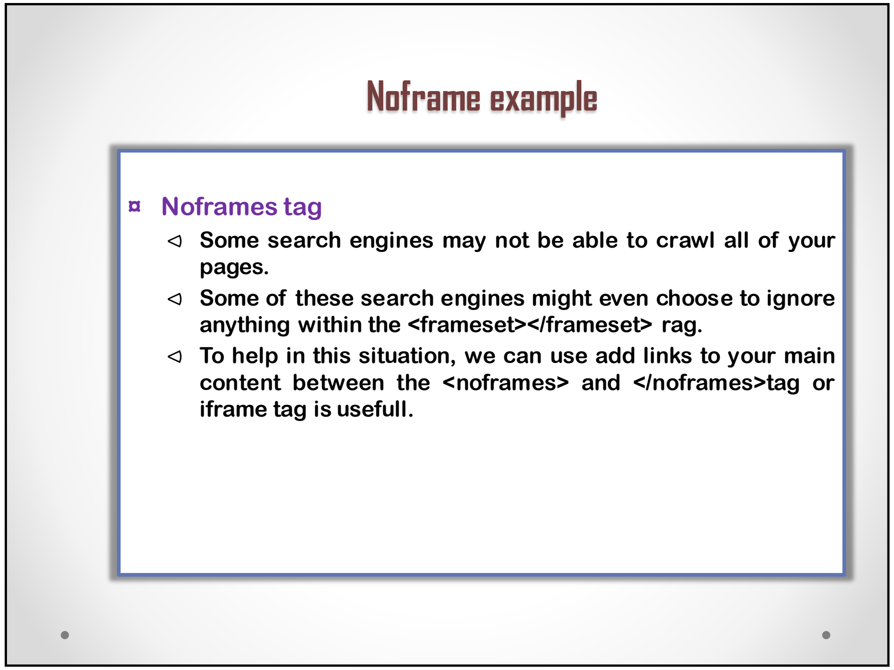
	Content
Navigation	

    - ✓ The navigation code appears below the content in the physical file.
    - ✓ But it still displays on the left when loaded in browser.



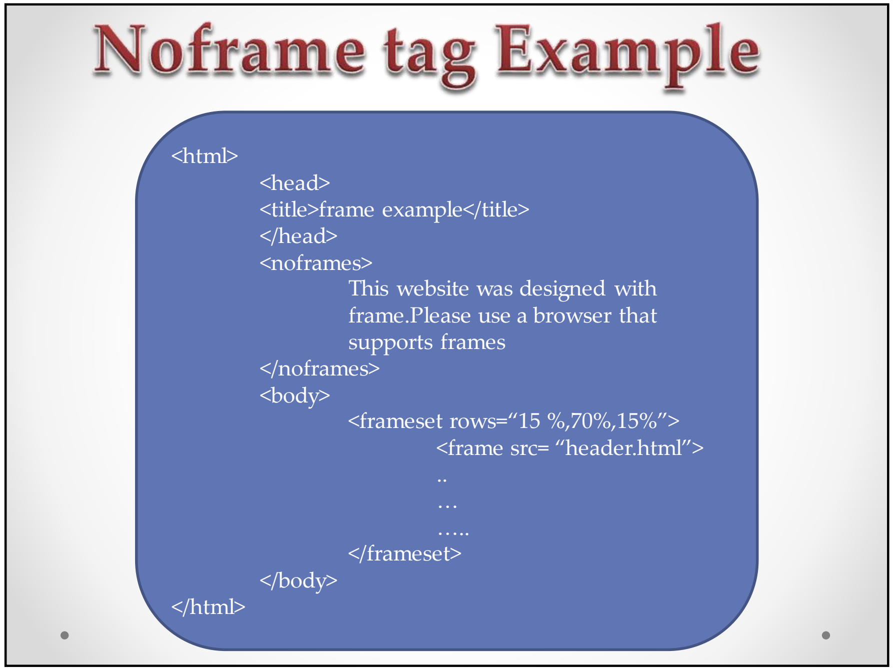
Search Engine – Friendly HTML

- Frames
  - Search engine have a lot of trouble spidering frames based sites.
  - Search engine can not index a frames page within the context of its other associated frames.
  - The noframes tag also attempts to address the problem.
  - But it is an invisible on-page factor and mercilessly abused by spammers.
  - It is suggestion that not to use such a frames.



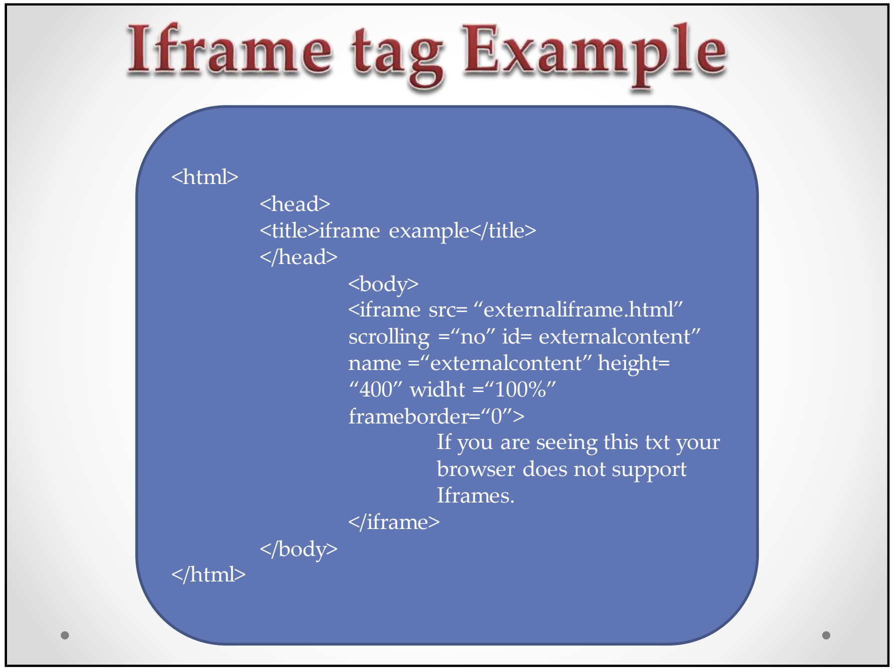
Noframe example

- Noframes tag
  - Some search engines may not be able to crawl all of your pages.
  - Some of these search engines might even choose to ignore anything within the <frameset></frameset> rag.
  - To help in this situation, we can use add links to your main content between the <noframes> and </noframes>tag or iframe tag is usefull.



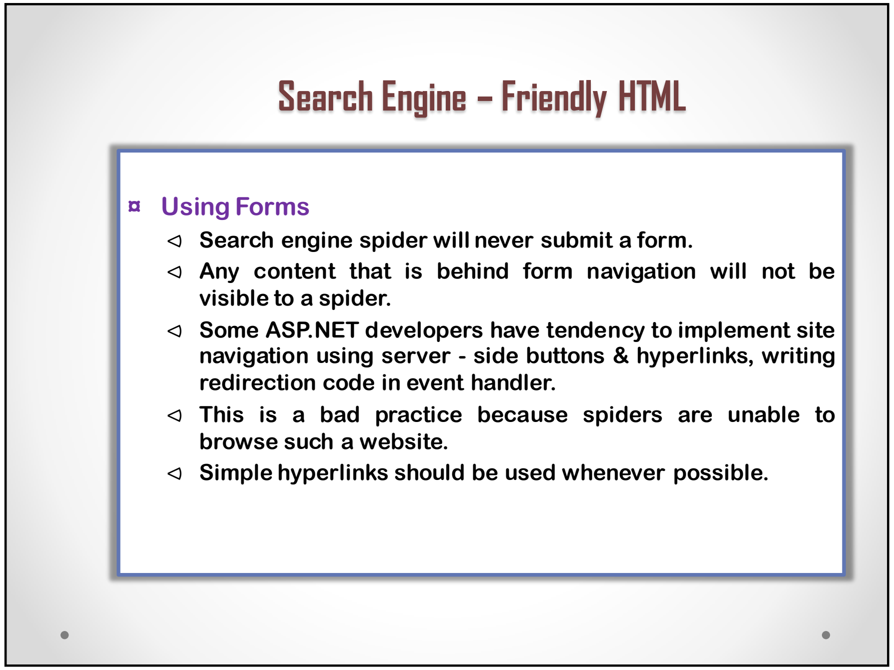
Noframe tag Example

```
<html>
        <head>
        <title>frame example</title>
        </head>
        <noframes>
                This website was designed with
                frame.Please use a browser that
                supports frames
        </noframes>
        <body>
                <frameset rows="15 %,70%,15%">
                        <frame src= "header.html">
                        ..
                        …
                        …..
                </frameset>
        </body>
</html>
```



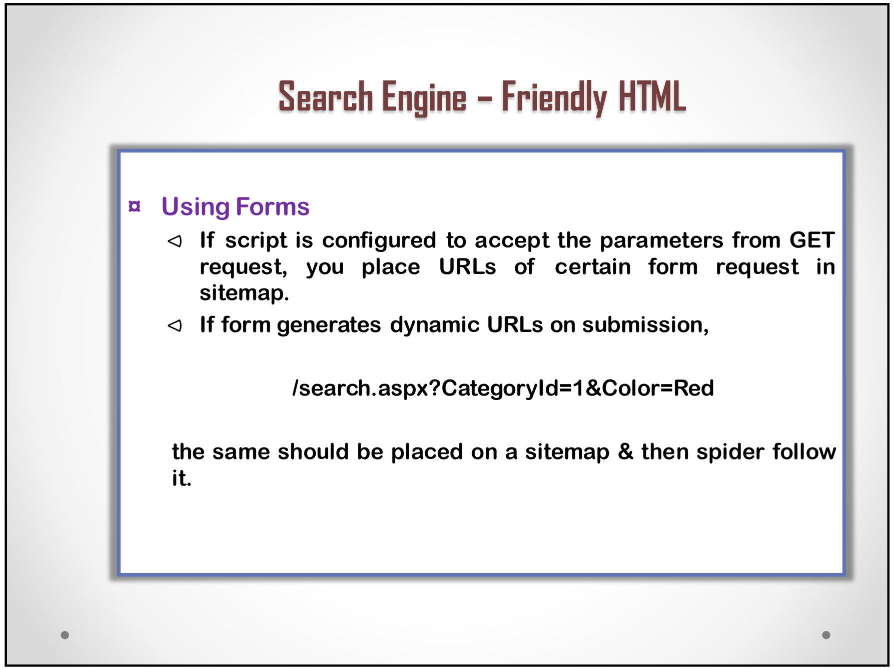
Iframe tag Example

```
<html>
        <head>
        <title>iframe example</title>
        </head>
                <body>
                <iframe src= "externaliframe.html"
                scrolling ="no" id= externalcontent"
                name ="externalcontent" height=
                "400" widht ="100%"
                frameborder="0">
                        If you are seeing this txt your
                        browser does not support
                        Iframes.
                </iframe>
        </body>
</html>
```



Search Engine – Friendly HTML

- ¤ Using Forms
  - ◁ Search engine spider will never submit a form.
  - ◁ Any content that is behind form navigation will not be visible to a spider.
  - ◁ Some ASP.NET developers have tendency to implement site navigation using server - side buttons & hyperlinks, writing redirection code in event handler.
  - ◁ This is a bad practice because spiders are unable to browse such a website.
  - ◁ Simple hyperlinks should be used whenever possible.



Search Engine – Friendly HTML

- ¤ Using Forms
  - ◁ If script is configured to accept the parameters from GET request, you place URLs of certain form request in sitemap.
  - ◁ If form generates dynamic URLs on submission,

/search.aspx?CategoryId=1&Color=Red

the same should be placed on a sitemap & then spider follow it.