Search Engines Series Pt. 2 - Coding The Header

Posted At : December 13, 2006 12:44 PM | Posted By : Mark Kruger
Related Categories: Hosting and Networking

In this post, part 2 in our search engine series, we will discuss important aspects of coding and designing that will facilitate easier indexing by search engines and create a higher likelihood of a rising page rank. In **Part 1** of this series we discussed the concept that your web site needs valuable and fresh content to really be useful to search engines. Without useful content your web site will not be a destination that anyone *wants* to visit, and therefore it will not be something that search engines (who are customer focused) want to index. Keep part 1 in mind as we discuss what you can do with your code and with your pages. Unless you have solved the puzzle of maintaining fresh and valuable content on your web site, you will be spinning your wheels.

Of course, you can use certain techniques to get yourself ranked high - at least temporarily. But my guess is that you will spend just as much time changing your code in a running battle to keep yourself on top of Google as you would if you learned to write and maintain good information on your site. By the way, those "black-hat techniques" are *not* discussed in this post. This post is about preparing your valuable content to be consumed by a search engine that wants and needs it - not about tricking a search engine into indexing less than worthy content. With that in mind....

The Header

You remember the header? It's that thing above the tag? It has some important items in it that are relevant to search engines. Here's the header from part 1 of this series:

```
<!DOCTYPE HTML
    PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
    "http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
    <title>Coldfusion Muse: Search Engines Series Pt. 1 - Content is King</title>

    <InvalidTag name="robots" content="index,follow" />
    <InvalidTag name="title" content="Coldfusion Muse: Search Engines Series Pt. 1 - Content
is King" />
    <InvalidTag content="text/html; charset=UTF-8" http-equiv="content-type" />
    <InvalidTag name="description"
        content="Mark Kruger's Daily Musings on Coldfusion: Search Engines Series Pt. 1 -
Content is King" />
    <InvalidTag name="keywords"
        content="Coldfusion,CFMX,..." />

    <link rel="stylesheet" type="text/css"
            href="/r/s/aura_screen_layout.css"
            media="screen" />
        .... plus a bunch more style info....

    <link rel="alternate" type="application/rss+xml"
        title="RSS" href="/rss.cfm?mode=full" />

    <InvalidTag type="text/javascript">
            ... function for popups...
    </script>
</head>
```

Now lest I take credit for this header I will hasten to say it is not my creation. I'm using Ray Camden's **BlogCFC** version 5 (I know I'm a version behind). This blog has always been indexed frequently and it shows up nicely on searches related to Coldfusion, so I'm not changing anything Ray did ('cause he's a pretty smart guy). Instead let's peruse some of the features of this header and make some assessments of what each thing does and how important it is.

Document Declaration (DOCTYPE)

Let me say at the outset that it is important to *have* a document declaration. Your browser attempts to use this declaration to render your page. The "DTD" for the document declaration determines "rules" for how the entities in your page are displayed. In our case we are using "4.01 transitional" with the "loose" dtd. That means the browser will not try to "strictly enforce" the rules. It will make it's "best effort" to render the page. To see how this works, take an older page with lots of tables and inline styles and change the Document declaration to strict:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
        "http://www.w3.org/TR/html4/strict.dtd">
```

You will find that the page looks very different in most cases. Why? Because the rules for displaying a strict "HTML" page are different from displaying a transitional HTML page.

What does that have to do with search engines? The rule of thumb is that you *should have* a document declaration and that your page should conform to that declaration. Does that mean search engines will validate your HTML or XHTML to see if it meets the requirements? That is highly unlikely, and not just because it would be a vast undertaking. The truth is that almost no web sites validate correctly. As a test I ran some sites through the **W3 Validator** with some surprising results:

- **www.yahoo.com** - No Document declaration. 43 validation errors (keep in mind the vaidator will use 4.01 transitional when no doc type is specified).
- **www.google.com** - No doc declaration, 63 errors.
- **www.msn.com** - Validation passed with an XHTML "strict" declaration.
- **www.ebay.com** - No declaration, 225 errors.
- **www.adobe.com** - XMTML transitional declaration, 14 errors.

A couple things "jump out". One is that the only site to pass the standard's test is owned by Microsoft. That makes me think that the lack of validation everywhere else might be just a knee-jerk reaction (If Microsoft's following the standard then by-jimminy we ain't a gonna do it). I'm kidding of course, but I wouldn't be surprised If someone believed it. The other thing that jumps out is that many large sites with boat loads of content have *no* document declaration. Of course sites like yahoo and Google are *indexers* not *indexees*. In any case, the fact that adobe comes in with only 14 errors using "XHTML transitional" is reassuring. Bottom line with doctype and search engines – make sure you have one, but don't hang your hat on it.

Title, Keywords and Descriptions

One of my developers, Jason Herbolsheimer, came back from MAX having attended a search engine workshop with **Figleaf's** own Steve Drucker. Jason's take on Steve's session was that the *title* is possibly the most important item on any given page for

search engine ranking. In other words, every page should have a title and the title should reflect the content on the page. In this respect Ray's blog does a nice job for me by putting the title of the story into the title tag. It also puts it into a meta tag named "title", and the title (with slight variation) goes into the meta tag "description" as well.

Regarding those famously overused and popular meta tags of "keyword" and "description" - they are much less widely used than you might think. In fact, according to this article on **meta tags** in Wikipedia (last updated Nov 16, 2006), Google *doesn't trust* and *doesn't use* meta data for indexing although the article does say that some researchers have found a link between visibility and page ranking. My advice? Make sure you have the "description" and "keyword" meta tags, but don't put your faith in them. Instead, pay attention to the title tag!

Adding "Other Stuff" to the Header

The header above is nice and clean. It includes the things a header should include. You might notice what it does *not* include. It does not include a style block and it does not include *excessive* JavaScript. I have seen (and developed) many pages with a long style block in the header like this:

```
<STYLE type=text/css>
    ul {....    }
    ul a { font: bold 110%; }
    ul li { margin: 4px 0px; padding: 0px 0px 0px 0px; }
    .submit {    ...     }
.nav {...}
.nav:hover {...}
.nav2 {...}
.Headline {...}
.Headlinecenter    { ...}
    .Headlinecenter td {....}
    .... more classes and styles ....
</STYLE>
```

Or perhaps a list of JS functions that are needed (or may not be needed) in the page. These items are candidates for external linking. Unless you have just a couple of styles to include, put your styles in a .css style sheet and link to it (like the example). Unless you have more than a couple simple functions (like a popup function) put your JS in an external file and load it with the src attribute. Don't clutter up the header with lots of "non Search Engine Worthy" stuff. You don't want the search engine to have to "figure out" what is useful and what can be ignored. Keep it simple and clean.

RSS

While we are on the subject of simple and clean, nothing is cleaner or simpler or easier to index than RSS. As search engines get more and more cluttered with the waste by-products of black-hat optimization, blogs and RSS are the "new" way to get to useful content. In my opinion, you should not consider building a traffic driven site without giving some thought to RSS as a supplement. Even if a search engine does not index your RSS you should remember that many sites are consuming RSS feeds. With RSS you increase the likelihood that some site *will link* to your site (as long as you have good content) - and that is definitely a marker that search engines take into account. If other sites link to yours, your page ranking will rise.

The Robot Meta Tag

This little tag can be used to instruct a search engine spider what to do with your page. You have 3 indicators to work with, index, follow, and no.

- index - tells the search engine to index *this page*.
- follow - tells the search engine to *follow links* from this page to subsequent pages (it tells it to spider).
- no - tells it to not index or not follow. The word “no” is appended to the other 2 with no spaces (as in noindex and nofollow)

Given these parameters you could tell a search engine to *index this page and follow links to subsequent pages*.

```
<InvalidTag name="robots" content="index,follow" />
```

...don't index this page but follow links to subsequent pages...

```
<InvalidTag name="robots" content="noindex,follow" />
```

...index this page but do not follow links from this page....

```
<InvalidTag name="robots" content="index,nofollow" />
```

...don't do anything with this page or any subsequent pages...

```
<InvalidTag name="robots" content="noindex,nofollow" />
```

It should be noted in regard to the meta "robot" tag and the "robots.txt" that more than one blog or webmaster resource I ran across put little faith in either of these methods. They indicated that it is their belief that spiders and agents simply ignore these instructions and index everything that they can get their hands on. One other note in regard to the "nofollow" tag, don't confuse the meta tag for robots or the robots.txt file with the *rel="nofollow"* that you sometimes see on a link. The rel="nofollow" attribute is to tell search engines not to include this link in its ranking. Primarily this is used to cut down on the efficacy of link spammers. See this Wikipedia article on **Spam in Blogs** for a good explanation of the ins and outs of this technique.

More to Come

In our next post we will continue the discussion on coding techniques and delve into how your page is constructed.