

[Home](#)

Main Products:

- [FDSE](#)
- [AXS](#)
- [Guardian](#)

[Forum](#)

[Contact Info](#)

[About](#)

[Site Map](#)

©2006 Zoltan Milosevic.
All Rights Reserved.

[Home](#) > [Fluid Dynamics Search Engine](#) > [Help](#) > **1048**

How to prevent sections of your pages from being indexed

This help topic describes how to prevent *sections* of a document from being indexed. To prevent an *entire* document from being indexed, see [How to prevent your pages from being indexed](#).

FDSE supports the proprietary "robots" comment tag. This tag allows a web author to apply robots exclusion rules to arbitrary sections of a document. The tag has one attribute, content, with the following possible values:

- noindex - the text enclosed in the tag is not saved in the index
- nofollow - links are not extracted from the text enclosed
- none - enclosed text is not indexed nor searched for links

Values "index", "follow", and "all" are also valid. In practice they are ignored since they are the unspoken defaults.

This feature is expected to fit the customer need of preventing certain parts of a document - such as a navigational sidebar - from being included in the search.

Example:

```
<HTML>
<BODY>

  This text will be indexed.
  <a href="foo.html"> this link will be followed </A>

  <!-- robots content="none" -->

    This text will NOT be indexed.
    <a href="bar.html"> this link will NOT be followed </A>

  <!-- /robots -->

  <!-- robots content="noindex" -->

    This text will NOT be indexed.
    <a href="bar1.html"> this link WILL be followed </A>

  <!-- /robots -->

  <!-- robots content="nofollow" -->

    This text WILL be indexed.
    <a href="bar1.html"> this link will NOT be followed </A>

  <!-- /robots -->

  la la la

</BODY>
</HTML>
```

For the example of a navigational sidebar, the "noindex" vale would be the best choice.

This syntax was designed to match the robots META tag.

For documents which have both the "robots" META tag and the "robots" comment tag, the most restrictive interpretation will be made, always erring on the side on not indexing or not following.

This tag is *always* respected by FDSE, even when "Crawler: Rogue" is set to 1 and all other robots exclusion rules are ignored. This is because the robots comment tag is proprietary and thus is expected to only be used by the FDSE administrator.

Coding:

All search engines are encouraged to support this syntax, since it results in small index files, faster searches, and more accurate results, and a standard syntax for authors to use. Currently Darryl Burgdorf's [WebSearch](#) supports this syntax too. I've asked Atomz, Google, and Altavista to support it and have received some favorable feedback but I'm not aware of any code changes on their part.

In Perl, the following regular expression will extract the robots sections labeled "none". This regex is designed to be insensitive to whitespace, line breaks, case, and the quoted-vs-bare nature of the attribute value. This regex should be executed as soon as the file is read:

```
$_ = $text;
s'<!--\s*robots\s+content\s*=\s*"?none"? \s*-->.*?<!--\s*/robots\s*-->' 'isg;
```

For subroutines which extract links, this code will extract the "nofollow" sections:

```
s'<!--\s*robots\s+content\s*=\s*"?nofollow"? \s*-->.*?<!--\s*/robots\s*-->' 'isg;
```

A final substitution against the "noindex" portion of the file is required before indexing or searching for keywords. For those search engines which just read in text and index it, without extracting links, the more efficient single regex will work:

```
s'<!--\s*robots\s+content\s*=\s*"?(none|noindex)"? \s*-->.*?<!--\s*/robots\s*-->' 'isg;
```

See also [Compatibility with Atomz.com noindex tag](#).

History: originally introduced as the FDSE:ROBOTS HTML tag in version 2.0.0.0031.

As of version 2.0.0.0044, the HTML comment equivalent was introduced. This equivalent format is preferred because it will not cause problems with editors that only support strictly valid HTML, and it will not cause problems with nesting, and it can be adopted as a vendor-independent standard. The equivalent forms are:

```
<FDSE:ROBOTS value="none"> xxx </FDSE:ROBOTS>
<!-- robots content="none" --> xxx <!-- /robots -->
```

```
<FDSE:ROBOTS value="nofollow"> xxx </FDSE:ROBOTS>
<!-- robots content="nofollow" --> xxx <!-- /robots -->

<FDSE:ROBOTS value="noindex"> xxx </FDSE:ROBOTS>
<!-- robots content="noindex" --> xxx <!-- /robots -->
```

Handling of these tags is case insensitive and is insensitive to linear and vertical whitespace.

Bugs / reverse compatibility: in early versions, FDSE used the singular format "FDSE:ROBOT" and "!-- robot content". It was later realized that for compatibility with the "robots" META tag and the "robots.txt" file, the plural format would have to be used. Thus, as of version 2.0.0.0048, the FDSE software supports both singular and plural forms, and the documentation promotes the plural form. If you have an earlier version of FDSE software, then you should use the singular form "robot" rather than "robots" in creating your tags. Sorry.

"How to prevent sections of your pages from being indexed"
<http://www.xav.com/scripts/search/help/1048.html>