Wt - Bug #7401

removal of '<meta name="robots" content="noindex, nofollow">'

01/14/2020 10:46 PM - Ray .

Status:	New	**Start date:**	01/14/2020
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			

Description

Due to the single page nature of Wt, I can understand why this header is inserted into the page.

<meta name="robots" content="noindex, nofollow">

However, when invoking

removeMetaHeader(Wt::MetaHeaderType::Meta, "robots");

in the constructor or in initialize(), the header is not removed.

History

#1 - 01/14/2020 10:58 PM - Ray .

Bug #6216 covers this in a fashion, but there are always new bots and scanners coming and going.

And my configuration uses the progressive scan capability, so the page renders properly on the first pass of any inbound request.

I will play with the meta-headers in the config file to see if I can make a general default.

#2 - 01/14/2020 11:13 PM - Ray .

Going into the configuration file and using the following:

<head-matter user-agent=".*">

<meta name="robots" content="index,follow" />

</head-matter>

The header is seen twice:

<head>

<meta name="robots" content="index,follow"/>

<meta name="robots" content="noindex, nofollow">

#3 - 01/15/2020 08:49 PM - Ray .

My solution was to change the following to match my preferences and rebuild the project:

- src/web/skeleton/Boot.html
- src/web/skeleton/Hybrid.html

#4 - 01/16/2020 09:24 AM - Roel Standaert

Yes, it's indeed verbatim in those template files, so you can't remove them in code.

I think a big thing is that you don't want bots to index your pages with extra junk (session ids) in the URL, and they might just do that if you do it this way.

#5 - 01/16/2020 05:27 PM - Ray .

There are some bots which don't identify themselves as bots, but as multi-personality agents:

- Mozilla/5.0(Linux;U;Android 5.1.1;zh-CN;OPPO A33 Build/LMY47V) AppleWebKit/537.36(KHTML,like Gecko) Version/4.0 Chrome/40.0.2214.89 UCBrowser/11.7.0.953 Mobile Safari/537.36

- Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36
- Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.84 Safari/537.36

Most of those are up to no good, based upon the urls they are requesting. But there are probably some legitimate bots in there as well. But you probably know that already.

Could some way be designed for Wt::WApplication constructor or initialize() to signal the underlying server that the session is interactive or a bot, based upon heuristics at the application level?

And because some user agents do turn off javascript, there will probably be some user sessions which do have the 'extra junk' in the URL. So I don't know if segregation of duties is as clear cut as we would like.

The official party line is that Wt is for embedded systems, but I have yet to find any alternatives which provide similar functionality for general web purposes.

As such, would it be possible to add to the API as an alternative to initialize(), where page composition could be started (which starts some async processes to business logic and such), returns the thread to asio, and then provide a call/post to finish the initial page generation. In the current regime, I find I am locking a thread for 50 to 100 ms, which isn't very scalable. Or should I put this thought into a new bug id?