CUNY Academic Commons - Feature #3662

Duplicate Content/SEO/Google issues

2014-11-17 01:50 AM - Matt Gold

Status:	Assigned	**Start date:**	2014-11-17
Priority name:	Normal	**Due date:**	
Assignee:	Raymond Hoh	**% Done:**	0%
Category name:	SEO	**Estimated time:**	0.00 hour
Target version:	Future release		

Description

Hi All,

Those of you in the CAT group will have seen Bruce Rosenbloom's report on Google's changed search algorithms. It looks like they have serious consequences for the Commons and its users -- go to google, type CUNY Academic Commons, and the CAC doesn't even come up on the first page of results.

I checked in with Chris Caruso on this, and he suggested that the issue could be WordPress Multisite and duplicate content. You can see some of his thoughts on a previous SEO-related ticket here

I think that we need to try to start addressing some of these issues, as they will affect users who blog on the site or who otherwise hope to use it to raise their profiles by sharing their work on it.

Related issues:

Related to CUNY Academic Commons - Bug #2959: CUNY Academic Commons SEO issues	**Resolved**	**2014-01-14**

History

#1 - 2014-11-17 01:59 AM - Matt Gold

- File Advanced Wordpress SEO.pdf added

Here's a powerpoint that may have some useful advice around duplicate content on WP installs

#2 - 2014-11-17 08:18 PM - Boone Gorges

Thanks, Matt. I'm not sure how much of this is duplicate content, etc - Google does not make it clear. Do you have any specific information about Google's recent algorithm changes?

I just set up Google Webmaster Tools for the Commons, and it appears that we have a fairly large amount of uncrawlable links. I'm going to go through and start correcting them where possible. I'll also do something to prevent the crawler from going through our activity pagination, which I'm guessing from the GWT report has something to do with duplicate content reports.

As for the issue of duplicate content on subsites, I'm not sure there's much we can/should do. I'd be pretty wary of, say, changing everyone's themes to remove full post content from the home page and from category/tag/author archives. We can do some more experimentation with robots.txt across these sites.

Ray, do you have any thoughts about BuddyPress-specific stuff?

#3 - 2014-11-17 08:19 PM - Matt Gold

Thanks so much, Boone -- this sounds like a great start.

#4 - 2014-11-17 08:30 PM - Boone Gorges

Fatal error in MediaWiki RSS extension that was causing 500 errors on a number of pages fixed in https://github.com/cuny-academic-commons/cac/commit/907ff82aaeb409170169af12a61c012d84a7de1d

#5 - 2014-11-17 09:36 PM - Boone Gorges

- Target version set to 1.7.4

I've been playing with this for the last few hours, and I have an update. It turns out that Google basically stopped being able to index most pages on the Commons starting around 11/1/2014. I believe this is the direct cause of the depressed ranking in search results. GWT is telling me that it's been getting lots of 401 Forbidden errors. I've looked through every line of code that was deployed on 11/1, but I can't find anything obvious that would cause this kind of issue, so I'm going to guess that it's a coincidence that it happened on the same day as a release. Through some experimentation, I found out that disabling blocks in .htaccess that enable the HTTP authentication on cdev enabled the googlebot to reach the Commons. So, for the time being, I've commented out those lines on the production site. I'll wait a day or two to see if this kickstarts the indexing again, and if so, I'll figure out what can be done to make these changes permanent.

#6 - 2014-11-17 09:41 PM - Dominic Giglio

Boone,

I know this is a Ruby/Rails gem but we've had a lot of success using it to dynamically generate a sitemap that is sent to Google/Bing from a cron job once a week:

https://github.com/kjvarga/sitemap_generator

It lets us tell search engines what they should or should not be crawling. I assume there must be something similar for WP. I seem to remember Yoast's SEO plugin doing something like this:

https://yoast.com/wordpress/plugins/seo/

#7 - 2014-11-18 12:25 AM - Raymond Hoh

> Ray, do you have any thoughts about BuddyPress-specific stuff?

All of our BP user profile pages use the same <title> tag. (eg. USER's Profile| CUNY Academic Commons.) This is bad for SEO. (BP 2.1 doesn't suffer from this bug; the trunk version does though.)

I'm not sure whereabouts in the CAC codebase that the USER's Profile <title> override is occurring. Spent about 15 minutes trying to pinpoint where this is done, but couldn't find where.

#8 - 2014-11-18 03:48 AM - Raymond Hoh

I've just had a chance to read through the PDF document.

It would be easy enough to add in noindex meta tags on all archive pages (tag, date, author pages) and paginated pages throughout the Commons. The PDF recommends leaving the traditional category pages alone.

#9 - 2014-11-21 11:07 AM - Boone Gorges

- Assignee changed from Boone Gorges to Raymond Hoh

- Target version changed from 1.7.4 to 1.8

After much debugging, it looks like we've narrowed down the immediate problem to an incorrect server configuration, related to the recent rollout of a Varnish proxy. Google now appears to be in the process of reindexing the site, and we're back on top of our own Google search :)
https://www.google.com/?gws_rd=ssl#q=cuny+academic+commons

I'm going to put this ticket into the 1.8 release so we can look further into some of the improvements Ray suggests.

> I'm not sure whereabouts in the CAC codebase that the USER's Profile <title> override is occurring. Spent about 15 minutes trying to pinpoint where this is done, but couldn't find where.

Sorry about this - I think it's wp-content/themes/bp-nelo/cacap/home.php.

> It would be easy enough to add in noindex meta tags on all archive pages (tag, date, author pages) and paginated pages throughout the Commons. The PDF recommends leaving the traditional category pages alone.

I don't have a problem with doing this. Ray, can you write up a patch?

#10 - 2015-04-13 04:37 PM - Boone Gorges

- Target version changed from 1.8 to Future release

I'm removing this from the 1.8 milestone, as I think it's dependent on some BP ticket or other. Ray, do you think we should close this ticket altogether?

Files

Advanced Wordpress SEO.pdf	4.66 MB	2014-11-17	Matt Gold