Redmine - Defect #6734
robots.txt: disallow crawling issues list with a query string

2010-10-24 16:47 - Be Fio

Status:	Closed	**Start date:**	2010-10-24
Priority:	Normal	**Due date:**	
Assignee:	Go MAEDA	**% Done:**	0%
Category:	SEO	**Estimated time:**	0.00 hour
Target version:	4.0.8		
Resolution:	Fixed	**Affected version:**	1.0.2

Description

When robots visit robots.txt, it tells them to disallow /projects/project/issues, but nowhere does it tell it to disallow /issues

From looking at access logs, Googlebot (but all other bots do it to) was indexing /issues, and was indexing many useless pages, mainly like this:

> 66.249.68.115 - - [24/Oct/2010:07:05:00 -0700] "GET /issues?sort=assigned_to%2Cupdated_on%2Cstatus%3Adesc HTTP/1.1" 200 6254 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

There are about a few hundred of those entries. I disallowed the sort parameter with Google Webmaster Tools, but that's just working around the issue for now.

Related issues:

Related to Redmine - Defect # 7582: hiding form pages from search engines	**Closed**	**2011-02-09**
Related to Redmine - Patch # 3754: add some additional URL paths to robots.txt	**New**	**2009-08-18**
Related to Redmine - Feature # 31617: robots.txt: disallow crawling dynamical...	**Closed**	

Associated revisions

Revision 19844 - 2020-07-02 05:02 - Go MAEDA

robots.txt: disallow crawling issues list with a query string (#6734).

Patch by Go MAEDA.

Revision 19852 - 2020-07-04 02:52 - Go MAEDA

Merged r19844 from trunk to 4.1-stable (#6734).

Revision 19853 - 2020-07-04 02:53 - Go MAEDA

Merged r19844 from trunk to 4.0-stable (#6734).

History

#1 - 2010-10-24 16:49 - Be Fio

Aww, excuse me for putting this in "search engine", just realized the category doesn't actually fit this but report. >.<

#2 - 2010-10-24 17:10 - Be Fio

From searching the Google index, it also appears that they have not indexed /projects/project/issues, but they did index /projects/project/issues?tracker_id=1, whether Googlebot is following the robots.txt mostly but not completely I do not know, but that page is indexed regardless, where it shouldn't be.

#3 - 2010-10-24 18:16 - Be Fio

You can just ignore comment 2.

#4 - 2010-10-25 12:39 - Felix Schäfer

- Category deleted (Search engine)

Do you have any idea/example on how to disable bots to navigate parametrized URLs?

#5 - 2010-10-25 20:15 - Be Fio

Hi,

From what documentation I can get my hands on, this doesn't seem to be documented. I know that putting an entry like:
Disallow: /issues

Will work, however I am guessing that not disallowing that might have been intentional.

After searching a bit however, I came across a bit of code that is said to work, but I haven't been able to verify it yet.

```
Disallow: *sort=
Disallow: *&sort=
Disallow: *? // This should disallow all URL's that request something, not necessarily a good idea, but it's just an example
Disallow: *sort=*
// if above's won't work, I heard that wildcards aren't supported, so maybe something like..
Disallow: /issues?sort=
```

I'm 75% sure the ones with the wildcards will work, and 90% sure the example without the wildcard will work.

I tried to put in as many examples as I could. Like I said, I couldn't and am unable to verify them though. Also, there may be more parameters that should be disallowed, but I missed (or they weren't yet navigated). I'll keep on the lookout for more, and update this report as needed. Hope that helps!

Please note: As you can see when visiting Redmine's robots.txt, it states some URL's to disallow. It appears that Googlebot disregards a lot of these even though it knows they're disallowed. I know this, because using Google Webmaster Tools, it showed me that the bot knows that they're disallowed URL's, even though it visited them.

#6 - 2010-10-25 23:43 - Felix Schäfer

Be Fio wrote:

> *From what documentation I can get my hands on, this doesn't seem to be documented. I know that putting an entry like:*
> *Disallow: /issues*
>
> *Will work, however I am guessing that not disallowing that might have been intentional.*

I guess so too, the one rule about the issue list is to prevent bots indexing stuff twice.

> *After searching a bit however, I came across a bit of code that is said to work, but I haven't been able to verify it yet.*
> *[...]*
>
> *I'm 75% sure the ones with the wildcards will work, and 90% sure the example without the wildcard will work.*

So sort of "official" documentation would be nice, or at least confirmation that this works. Care to share your sources?

> *I tried to put in as many examples as I could. Like I said, I couldn't and am unable to verify them though. Also, there may be more parameters that should be disallowed, but I missed (or they weren't yet navigated). I'll keep on the lookout for more, and update this report as needed. Hope that helps!*
>
> *Please note: As you can see when visiting Redmine's robots.txt, it states some URL's to disallow. It appears that Googlebot disregards a lot of these even though it knows they're disallowed. I know this, because using Google Webmaster Tools, it showed me that the bot knows that they're disallowed URL's, even though it visited them.*

That's a problem you should tackle with google, not with us ;-)

#7 - 2010-10-26 09:56 - Be Fio

This isn't "official", but I tried to find as many sources as I could, hopefully these'll help out. From what I read, the wildcards will work (for some bots), but are a bad idea because others won't follow it.
http://www.webmasterworld.com/forum93/823.htm
http://www.ihelpyou.com/forums/showthread.php?t=27849
http://www.velocityreviews.com/forums/t608728-robots-txt-and-regular-expressions.html

This is already following the standards, so it's a safe fallback:

```
Disallow: /issues?sort=
```

Conclusion: Wildcards are too risky, but we of course already know that the above will work normally as it conforms to the rules. It's up to you if you want to do something, or nothing. ;)
Official documentation: http://www.robotstxt.org/robotstxt.html

Felix Schäfer wrote:

> *That's a problem you should tackle with google, not with us ;-)*

Oh, I was just noting that so that you guys know about it. Like a "warning" :)

#8 - 2010-10-26 09:58 - Be Fio

Oh, and if /issues?sort= isn't the only one that bots might follow (because there's other parameterized stuff on the page), I suppose it'd probably be good to maybe put those in. I don't know all of the possible parameters, but you guys should. :)

#9 - 2011-01-12 04:33 - Be Fio

An alternative, and MUCH MUCH better solution, is to add a noindex meta tag to the pages that shouldn't be indexed, which there are a lot of those on Redmine that robots.txt doesn't cover, and Google is going crazy indexing them.

```
# tell robots not to index that page
<meta name="robots" content="noindex">

# this page is the same as this other page (good for when /issues/21/?reply=2 is the same as /issues/21/)
<link rel="canonical" href="/">
```

I highly suggest this gets implemented as soon as possible. :)

#10 - 2011-05-09 06:27 - Antoine Beaupré

It seems like this could be easily fixed by the patch is #3754.

#11 - 2019-06-24 15:59 - Harald Welte

We've just observed that this issue still exists in redmine 3.4. I couldn't find any rationale here in this issue why the related patch was not merged during the past 8 years at some point?

#12 - 2020-06-22 21:38 - Eduardo Ramos

Harald Welte wrote:

> *We've just observed that this issue still exists in redmine 3.4. I couldn't find any rationale here in this issue why the related patch was not merged during the past 8 years at some point?*

Still failing in redmine 4.1.1 stable, similar GETs on issues.
Receiving requests from various bots which exhaust my raspberry cpu:

172.162.119.114.in-addr.arpa domain name pointer petalbot-114-119-162-172.aspiegel.com.
146.168.229.46.in-addr.arpa domain name pointer crawl18.bl.semrush.com.
...

#13 - 2020-06-23 04:17 - Go MAEDA

- Category set to SEO
- Target version set to Candidate for next minor release

Most people here think that the problem is that search engines indexes URLs of filters and queries (/issues/?...) rather than single issue pages (/issues/123).

I agree that indexing "/issues/?..." URLs is a waste of computer resources. However, I think "/issues/123" URLs should be indexed (I usually search for issues in www.redmine.org with Google).

The following patch disallows all URLs that have a query string (?...). It disallows indexing "/issues/?" pages while allowing indexing "/issues/123" pages. The main contents we want search engines to index are issues and wiki pages, so I think it is not a problem to disallow all URLs that have a query string.

```
diff --git a/app/views/welcome/robots.text.erb b/app/views/welcome/robots.text.erb
index 6f66278ad..dbe9f04dd 100644
--- a/app/views/welcome/robots.text.erb
+++ b/app/views/welcome/robots.text.erb
@@ -1,4 +1,5 @@
 User-agent: *
+Disallow: /*?
 <% @projects.each do |project| -%>
 <%   [project, project.id].each do |p| -%>
 Disallow: <%= url_for(:controller => 'repositories', :action => :show, :id => p) %>
```

#14 - 2020-06-23 14:45 - Eduardo Ramos

Go MAEDA wrote:

> *Most people here think that the problem is that search engines indexes URLs of filters and queries (/issues/?...) rather than single issue pages (/issues/123).*
>
> *I agree that indexing "/issues/?..." URLs is a waste of computer resources. However, I think "/issues/123" URLs should be indexed (I usually search for issues in www.redmine.org with Google).*
>
> *The following patch disallows all URLs that have a query string (?...). It disallows indexing "/issues/?" pages while allowing indexing "/issues/123" pages. The main contents we want search engines to index are issues and wiki pages, so I think it is not a problem to disallow all URLs that have a query string.*
>
> *[...]*

Thank u, with that modification (Disallow: /*?) my raspberry is not so stressed (at least cpu is under 15%, before it was about 90% due to crawlers requests)
How could it be patched on a docker-compose layout ? What I did, is to modify the 'robots.text.erb' in redmine container, and restart such container.

#15 - 2020-06-23 14:55 - Go MAEDA

- Target version changed from Candidate for next minor release to 4.0.8

Eduardo Ramos wrote:

> *Thank u, with that modification (Disallow: /*?) my raspberry is not so stressed (at least cpu is under 15%, before it was about 90% due to crawlers requests)*

Thank you for testing the patch and for giving feedback. I am setting the target version to 4.0.8.

> *How could it be patched on a docker-compose layout ? What I did, is to modify the 'robots.text.erb' in redmine container, and restart such container.*

I don't know much about Docker. I suggest you ask questions on the forums.

#16 - 2020-06-23 16:48 - Go MAEDA

Updated the patch. The previous patch posted in #6734#note-13 has a problem that it prevents crawlers from accessing "/issues?page=". It means that crawlers can get only the first page of the issues list and will not index issues after the second page.

```
diff --git a/app/views/welcome/robots.text.erb b/app/views/welcome/robots.text.erb
index 6f66278ad..8c2732e00 100644
--- a/app/views/welcome/robots.text.erb
+++ b/app/views/welcome/robots.text.erb
@@ -10,3 +10,6 @@ Disallow: <%= url_for(issues_gantt_path) %>
 Disallow: <%= url_for(issues_calendar_path) %>
 Disallow: <%= url_for(activity_path) %>
 Disallow: <%= url_for(search_path) %>
+Disallow: <%= url_for(issues_path) %>?sort=
+Disallow: <%= url_for(issues_path) %>?query_id=
+Disallow: <%= url_for(issues_path) %>?*set_filter=
```

#17 - 2020-06-23 17:03 - Go MAEDA

- Related to Feature #31617: robots.txt: disallow crawling dynamically generated PDF documents added

#18 - 2020-06-23 17:42 - Eduardo Ramos

Go MAEDA wrote:

> *Updated the patch. The previous patch posted in #6734#note-13 has a problem that it prevents crawlers from accessing "/issues?page=". It means that crawlers can get only the first page of the issues list and will not index issues after the second page.*
>
> *[...]*

Tested OK. The cpu even better. No activity registered in redmine logs regarding bots, neither at redmine access logs from nginx.
It could be casuality (no crawlers accessing now), i will monitor it in the following hours anyway.

Thank u

#19 - 2020-06-24 04:43 - Go MAEDA

- Subject changed from Robots index /issues (which isn't disallowed in robots.txt) to robots.txt: disallow crawling issues list with a query string

#20 - 2020-06-28 09:33 - Go MAEDA

- File 6734.diff added

Added test code.

#21 - 2020-06-28 09:39 - Go MAEDA

- File 6734.patch added

#22 - 2020-06-28 09:39 - Go MAEDA

- File deleted (6734.diff)

#23 - 2020-07-02 05:02 - Go MAEDA

- Status changed from New to Resolved

- Assignee set to Go MAEDA

- Resolution set to Fixed

Committed the patch.

#24 - 2020-07-04 02:54 - Go MAEDA

- Status changed from Resolved to Closed

Files

6734.patch	1.14 KB	2020-06-28	Go MAEDA