Redmine - Defect #27865

RailsBaseURI ignored while creating robots.txt

2017-12-28 20:27 - Grischa Zengel

Status:	Closed	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:	SEO	**Estimated time:**	0.00 hour
Target version:	4.0.0		
Resolution:	Fixed	**Affected version:**	3.4.3

Description

In my case I use
RailsBaseURI /redmine
which results to URLs like /redmine/issues but in robots.txt are URLs without the prefix /redmine:

```
User-agent: *
Disallow: /projects/support/repository
Disallow: /projects/support/issues
Disallow: /projects/support/activity
Disallow: /issues/gantt
Disallow: /issues/calendar
Disallow: /activity
Disallow: /search
```

The expected robots.txt must be

```
User-agent: *
Disallow: /redmine/projects/support/repository
Disallow: /redmine/projects/support/issues
Disallow: /redmine/projects/support/activity
Disallow: /redmine/issues/gantt
Disallow: /redmine/issues/calendar
Disallow: /redmine/activity
Disallow: /redmine/search
```

As a feature request, it would be nice to have some additional URLs which I could add in configuration.

Something like:

```
User-agent: *
<% @projects.each do |p| -%>
Disallow: /projects/<%= p.to_param %>/repository
Disallow: /projects/<%= p.to_param %>/issues
Disallow: /projects/<%= p.to_param %>/activity
<% @config.robots_project.each do |u| -%>
Disallow: /projects/<%= p.to_param %>/<%= u.to_param %>
<% end -%>
<% end -%>
Disallow: /issues/gantt
Disallow: /issues/calendar
Disallow: /activity
Disallow: /search
<% @config.robots_main.each do |u| -%>
Disallow: /<%= u.to_param %>
<% end -%>
```

Associated revisions

Revision 17135 - 2017-12-29 15:56 - Toshi MARUYAMA

use relative url at robots.txt (#27865)

Revision 17136 - 2017-12-29 16:09 - Toshi MARUYAMA

source code comment slight change (#27865)

History

#1 - 2017-12-28 21:55 - Toshi MARUYAMA

- File robots.diff added

- Status changed from New to Confirmed

- Target version set to 4.0.0

Try this patch.

#2 - 2017-12-28 22:28 - Grischa Zengel

I tested your patch and it works like expected.
It's clever to use the existing functions instead of assembling some string.
Thanks.

I got:

```
User-agent: *
Disallow: /redmine/projects/support/repository
Disallow: /redmine/projects/support/issues
Disallow: /redmine/projects/support/activity
Disallow: /redmine/issues/gantt
Disallow: /redmine/issues/calendar
Disallow: /redmine/activity
Disallow: /redmine/search
```

#3 - 2017-12-29 13:08 - Toshi MARUYAMA

- Status changed from Confirmed to New

- Target version deleted (4.0.0)

https://webmasters.stackexchange.com/questions/89395/robots-txt-should-be-in-the-root-directory-or-can-be-in-sub-directory

> web crawlers will not read or obey a robots.txt file in a subdirectory.
> http://www.robotstxt.org/robotstxt.html

#4 - 2017-12-29 13:24 - Grischa Zengel

Sure?

> RedirectMatch permanent ^/robots.txt$ /redmine/robots.txt

#5 - 2017-12-29 13:25 - Grischa Zengel

I would like to post my apache.log which show that all bots redirect properly, but I've been blocked always because of spam.

#6 - 2017-12-29 13:29 - Grischa Zengel

```
redmine.:443 88.198.55.175 - - [29/Dec/2017:08:10:23 +0000] "GET /robots.txt HTTP/1.1" 301 3962 "-" "Mozilla/5
.0 (compatible; MJ12bot/v1.4.8;)"
redmine.:443 88.198.55.175 - - [29/Dec/2017:08:10:23 +0000] "GET /redmine/robots.txt HTTP/1.1" 200 4126 "-" "M
ozilla/5.0 (compatible; MJ12bot/v1.4.8;)"
redmine.:443 66.249.76.118 - - [29/Dec/2017:09:48:48 +0000] "GET /robots.txt HTTP/1.1" 301 3962 "-" "Mozilla/5
.0 (compatible; Googlebot/2.1;)"
redmine.:443 66.249.76.118 - - [29/Dec/2017:09:48:48 +0000] "GET /redmine/robots.txt HTTP/1.1" 200 841 "-" "Mo
zilla/5.0 (compatible; Googlebot/2.1;)"
redmine.:443 5.255.251.125 - - [29/Dec/2017:10:51:21 +0000] "GET /robots.txt HTTP/1.1" 301 3967 "-" "Mozilla/5
.0 (compatible; YandexBot/3.0;)"
redmine.:443 87.250.233.120 - - [29/Dec/2017:10:51:21 +0000] "GET /redmine/robots.txt HTTP/1.1" 200 4187 "-" "
Mozilla/5.0 (compatible; YandexBot/3.0;)"
```

#7 - 2017-12-29 13:44 - Grischa Zengel

Some more:

```
redmine.:443 54.36.150.157 - - [28/Dec/2017:05:03:01 +0000] "GET /robots.txt HTTP/1.1" 301 3711 "-" "Mozilla/5.0 (compatible; AhrefsBot/5.2;)"
redmine.:443 54.36.150.157 - - [28/Dec/2017:05:03:02 +0000] "GET /redmine/robots.txt HTTP/1.1" 200 919 "-" "Mozilla/5.0 (compatible; AhrefsBot/5.2;)"
redmine.:443 157.55.39.25 - - [29/Dec/2017:01:35:11 +0000] "GET /robots.txt HTTP/1.1" 301 3942 "-" "Mozilla/5.0 (compatible; bingbot/2.0;)"
redmine.:443 157.55.39.25 - - [29/Dec/2017:01:35:12 +0000] "GET /redmine/robots.txt HTTP/1.1" 200 4012 "-" "Mozilla/5.0 (compatible; bingbot/2.0;)"
redmine.:443 194.187.170.123 - - [21/Dec/2017:22:26:10 +0000] "GET /robots.txt HTTP/1.0" 301 3825 "-" "Mozilla/5.0 (compatible; Qwantify/2.4w;)/2.4w"
redmine.:443 194.187.170.123 - - [21/Dec/2017:22:26:10 +0000] "GET /redmine/robots.txt HTTP/1.0" 200 1072 "-" "Mozilla/5.0 (compatible; Qwantify/2.4w;)/2.4w"
redmine.:443 18.195.89.56 - - [26/Dec/2017:00:37:01 +0000] "GET /robots.txt HTTP/1.1" 301 3711 "-" "Mozilla/5.0 (compatible; Cliqzbot/2.0;)"
redmine.:443 18.195.89.56 - - [26/Dec/2017:00:37:04 +0000] "GET /redmine/robots.txt HTTP/1.1" 200 782 "-" "Mozilla/5.0 (compatible; Cliqzbot/2.0;)"
```

#8 - 2017-12-29 13:51 - Toshi MARUYAMA

Grischa Zengel wrote:

> Sure?
>
> > RedirectMatch permanent ^/robots.txt$ /redmine/robots.txt

How will you do if your web site has plural subdirectories?

E.g.:

- example.com/redmine1
- example.com/redmine2

#9 - 2017-12-29 13:58 - Grischa Zengel

For better compatibility I will change to:

$ cat /etc/cron.hourly/robots

#!/bin/sh
wget https://redmine/redmine/robots.txt -O /var/www/html/robots.txt

#10 - 2017-12-29 14:01 - Grischa Zengel

> How will you do if your web site has plural subdirectories?

Than you have to use the cron solution and concatenate the results.
But how many servers will host more than one redmine instance? 1%?

#11 - 2017-12-29 14:16 - Toshi MARUYAMA

Grischa Zengel wrote:

> > How will you do if your web site has plural subdirectories?
>
> Than you have to use the cron solution and concatenate the results.

```
$ wget http://localhost:3100/test1/robots.txt -O - > ~/Desktop/robots.txt
$ wget http://localhost:3100/test2/robots.txt -O - | grep '^Disallow:' >> ~/Desktop/robots.txt
```

#12 - 2017-12-29 14:29 - Grischa Zengel

It doesn't matter how you post process the robots.txt from subdirectories, redmine has to generate a valid robots.txt and your patch works.
So please add it to next sub release, so I don't have to remember to patch manually. It doesn't break anything.

#13 - 2017-12-29 15:59 - Toshi MARUYAMA

- Status changed from New to Closed

- Target version set to 4.0.0

- Resolution set to Fixed

I have committed in r17135.

Grischa Zengel wrote:

> So please add it to next sub release

I don't want to change behaviour in sub version.

#14 - 2018-09-17 15:39 - Go MAEDA

- Category set to SEO

Files

robots.diff	857 Bytes	2017-12-28	Toshi MARUYAMA