Redmine - Feature #31617

robots.txt: disallow crawling dynamically generated PDF documents

2019-06-24 15:17 - Harald Welte

Status:	Closed	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	Go MAEDA	**% Done:**	0%
Category:	SEO	**Estimated time:**	0.00 hour
Target version:	4.2.0		
Resolution:	Fixed		

Description

While the auto-generated robots.txt contains URLS for /issues (the HTML issue list), it doesn't contain the same URLs for the PDF version.

At osmocom.org (where we use redmine), we're currently seeing lots of robot requests for /projects/*/issues.pdf?.... as well as /issues.pdf?....

Related issues:

Related to Redmine - Feature #3661: Configuration option to disable pdf creat...	**New**	**2009-07-23**
Related to Redmine - Defect #6734: robots.txt: disallow crawling issues list ...	**Closed**	**2010-10-24**

Associated revisions

Revision 19867 - 2020-07-09 02:33 - Go MAEDA

robots.txt: disallow crawling dynamically generated PDF documents (#31617).

Patch by Go MAEDA.

History

#1 - 2019-06-24 16:00 - Harald Welte

- Status changed from New to Resolved

I'm sorry, it seems the robot.txt standard is using sub-string matching, so foo/issues should include foo/issues.pdf. The crawler we see seems to be ignoring that :(

#2 - 2019-06-25 01:39 - Go MAEDA

- Category set to SEO

- Status changed from Resolved to Closed

- Resolution set to Invalid

Thank you for the feedback. Closing.

#3 - 2020-06-23 17:00 - Go MAEDA

- Status changed from Closed to Reopened

- Resolution deleted (Invalid)

The robots.txt generated by Redmine 4.1 does not disallow crawlers to access "/issues/<id>.pdf" and "/projects/<project_identifier>/wiki/<page_name>.pdf".

I think the following line should be added to the robots.txt.

```
Disallow: *.pdf
```

#4 - 2020-06-23 17:02 - Go MAEDA

- Related to Feature #3661: Configuration option to disable pdf creation of issues added

#5 - 2020-06-23 17:03 - Go MAEDA

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

#6 - 2020-06-24 02:25 - Go MAEDA

- Subject changed from robots.txt misses issues.pdf to robots.txt: disallow dynamically generated PDF

- Target version set to Candidate for next minor release

Since dynamically generated PDFs contain no more information than HTML pages and are useless for web surfers, the PDFs should not be indexed by search engines. In addition, In addition, generating a large number of PDFs in a short period of time is too much burden for a server.

I suggest disallowing web crawlers to fetch dynamically generated PDFs such as /projects/*/wiki/*.pdf and /issues/*.pdf by applying the following patch. The patch still allows crawlers to fetch static PDF files attached to issues or wiki pages (/attachments/*.pdf).

```
diff --git a/app/views/welcome/robots.text.erb b/app/views/welcome/robots.text.erb
index 6f66278ad..9cf7f39a6 100644
--- a/app/views/welcome/robots.text.erb
+++ b/app/views/welcome/robots.text.erb
@@ -10,3 +10,5 @@ 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(:trailing_slash => true)) %>*.pdf$
+Disallow: <%= url_for(projects_path(:trailing_slash => true)) %>*.pdf$
```

#7 - 2020-07-02 15:06 - Go MAEDA

- File 31617.patch added

- Subject changed from robots.txt: disallow dynamically generated PDF to robots.txt: disallow crawling dynamically generated PDF

- Target version changed from Candidate for next minor release to 4.2.0

Setting the target version to 4.2.0.

#8 - 2020-07-09 02:34 - Go MAEDA

- Tracker changed from Defect to Feature

- Subject changed from robots.txt: disallow crawling dynamically generated PDF to robots.txt: disallow crawling dynamically generated PDF documents

- Status changed from Reopened to Closed

- Assignee set to Go MAEDA

- Resolution set to Fixed

Committed the patch.

Files

31617.patch	1.16 KB	2020-07-02	Go MAEDA