Redmine - Feature #2646

Having a dynamic sitemaps file for search robots scanning

2009-02-02 14:04 - Axel Voitier

Status:	New	**Start date:**	2009-02-02
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			
Resolution:			

Description

I though it would be useful to have a controller that generate a sitemaps.xml file.
More info about sitemaps:
http://www.sitemaps.org/
http://en.wikipedia.org/wiki/Sitemaps

Such a thing would help indexation of projects, wikis, forums, and any other relevant pages for the most important bots.
Main benefit would be the ability to say to these bots when a page has been updated for the last time (especially for wikis pages).
It would go similarly to what have been recently done for the robots.txt file: #2491 and r2319.

In a more developed version it could allow the administrators (and maybe managers too) to set the periodicity scanning value for specifics pages (like news pages for instance, or, again, wikis pages), or give some pages a bigger importance value for indexing.

It is also imaginable to think about a "robots scan configuration tools" that fusion tunings for robots.txt and sitemaps.xml. Feature of such tool would be the ability to configure which pages should be scanned by (which) robots.

In the end: fine controls of what search robots can see.

History

#1 - 2009-02-02 16:23 - Axel Voitier

- File sitemaps.01.patch added

Here is a "kick off" patch for this feature.
It does not attend to be a useful one. It just start the work.

It does:

- Creation of a new controller and view named RobotsController and robots/sitemaps.rhtml
- List only wiki pages with their last update date in the sitemaps.xml file
- Add a route for that file

I guess the robots.txt generation could be added in this controlelr instead of Welcome.
This patch need modifications done in r2319! (mainly Project.public.active method).

TODOs:
- Check if a "http" or "https" have to be used following the Redmine configuration
- Add more pages (from projects, forums, etc.). Should be all pages publicly accessible in Redmine!

#2 - 2009-02-06 00:04 - Axel Voitier

Here is a list I made that lists every relevant pages a search engine should know about. It is ordered by controllers.

- account
  - /account/show/:id
- attachments
  - /attachments/:id/:filename
  - /attachments/download/:id/:filename
- boards

    - /projects/:project_id/boards
    - /projects/:project_id/boards/:id

- documents

    - /projects/:project_id/documents
    - /documents/:id

- issues

    - /issues
    - /projects/:project_id/issues
    - /issues/:id

- messages

    - /boards/:board_id/topics/:id

- news

    - /news
    - /projects/:project_id/news
    - /news/:id

- projects

    - /projects
    - /projects/:id/roadmap
    - /projects/:id/changelog
    - /projects/:id/files
    - /projects/:id/activity
    - /activity

- repositories

    - /projects/:id/repository/revisions
    - /projects/:id/repository/revisions/:rev
    - /projects/:id/repository/revisions/:rev/diff

- sys

    - /sys/service.wsdl

- welcome

    - /

- wiki

    - /projects/:id/wiki/:page
    - /projects/:id/wiki/Page_index
    - /projects/:id/wiki/Date_index

- wikis

    - /projects/:id/wiki

Does it needs some more pages? Or maybe less?

By the way, this lets me think that the robots.txt file disallowing pages for robots could be much more complete and fine detailed, in accordance with this list of relevant pages.

#3 - 2010-09-26 13:46 - Boris Pigeot

Thanks for your patch.
I think you don't need more pages.

For repository, I think it could be maybe too big.

Files

sitemaps.01.patch	1.51 KB	2009-02-02	Axel Voitier