TYPO3 Core - Feature #16534

Add possibility to start indexing an external site at a specific page

2006-09-06 20:13 - Mario Rimann

Status:	Closed	**Start date:**	2006-09-06
Priority:	Should have	**Due date:**	
Assignee:		**% Done:**	100%
Category:	Indexed Search	**Estimated time:**	0.00 hour
Target version:			
PHP Version:		**Complexity:**	easy
Tags:		**Sprint Focus:**	

Description

Current behaviour is that the starting URL is used for two purposes:
- determine where crawling starts
- check if the indexed pages are "inside" this URL

If you need to start the crawler at a specific page which is not a directory name there needs to be an extra setting.

Old description:
I noticed some strange behaviour when working with the indexed_search and the crawler extension: Some websites (like http://typo3.org/) are getting indexed including the subpages.

But on other domains, just the first page is indexed - but the links on that page are not followed (even if I configure it to dig 3 levels deep).

All the pages that aren't working are valid HTML or valid XHTML. I tried some different scenarios (like absolute/relativ paths as links) - no success.

TYPO3 4.0
Indexed search 2.9.0
Crawler 1.1.0
(issue imported from #M4167)

Associated revisions

Revision 819b5be0 - 2013-07-16 13:44 - Mario Rimann

[BUGFIX] Links on external pages don't get indexed

Allows the crawler to start indexing a specific file like
www.domain.tld/foobar.html instead of just www.domain.tld/

This is just about the comparison against the base URL and
enables the Crawler to start crawling at e.g. a file that contains
a manually generated list of links to follow. Before that change,
even links to targets on the same domain were rejected by
the checkUrl() method in case the base Url was pointing to some
file instead of "/". This was because the base URL was then not
part of the target URL.
After stripping off any path from the base URL for this comparison
this can now also be used to start crawling from a file.

Change-Id: I2727a9a447754b88d2c279c24b32b5c3a2df26c0
Resolves: #16534
Releases: 6.2, 6.1, 6.0, 4.7, 4.5
Reviewed-on: https://review.typo3.org/6990
Reviewed-by: Michael Stucki
Tested-by: Michael Stucki
Reviewed-by: Georg Ringer
Tested-by: Georg Ringer

History

#1 - 2006-09-09 21:09 - Mario Rimann

Some additional information:
I also tested this on another server running Debian Linux, Apache 1.3.something and TYPO3 v4.0.1. This test ended with the same result.

Tried to crawl some domains on the same server - some went well, other showed the same issue as described in bug report (first page fetched, but not followed the links on it).

I also noticed that it doesn't depend on whether HTML or XHTML is used. It also seems to be charset independent.

#2 - 2006-09-10 23:01 - Mario Rimann

More information:

I tracked this issue down to the function checkUrl() in the crawler class of indexed search.

If you start with a URL like "http://www.domain.tld/" (the root page):
- Links to inside of this domain will work
- Links to outside of that domain don't work

If you start with a URL like "http://www.domain.tld/fileadmin/linklist.htm" (a file / subfolder):
- No links will work! Neither absolute nor relative. As they get compared, it will never work out (the check will fail and the URL won't get added to the queue).

I think this should be enhanced by a configuration option to
a) ignore those checks and index "blind"
or
b) have a "whitelist" of domains (next to the BaseURL) and allow indexing for URLs start with that whitelist in any case.

I'd also appreciate if single files as base URL would be supported.

#3 - 2006-09-11 20:29 - Mario Rimann

I've attached an initial patch. This solves the problem if the baseURL (the starting URL for crawling/indexing) is pointing to a file instead of a domain-root.

This doesn't add any additional configuration options. Question to the CoreDevs: Which configuration options should be implemented? Any feedback is welcome!

#4 - 2009-08-28 10:58 - Mario Rimann

The second patch was adapted to the current trunk (rev. 5837)

#5 - 2009-08-28 12:05 - Ferdinand Kuhl

I read over the patch, and i looks very clean to me. It just allows the crawler to start with a file. All links inside the same domain as the file will be followed.

I tested it at a smaller test environment and it works for me.

#6 - 2010-10-05 12:50 - Dmitry Dulepov

I agree, the patch is good. We should get that into indexed search.

#7 - 2011-11-29 22:32 - Mr. Jenkins

- Status changed from Accepted to Under Review

Patch set 1 of change I2727a9a447754b88d2c279c24b32b5c3a2df26c0 has been pushed to the review server.
It is available at http://review.typo3.org/6990

#8 - 2011-11-30 22:55 - Mr. Jenkins

Patch set 2 of change I2727a9a447754b88d2c279c24b32b5c3a2df26c0 has been pushed to the review server.
It is available at http://review.typo3.org/6990

#9 - 2012-06-28 22:57 - Gerrit Code Review

Patch set 3 for branch **master** has been pushed to the review server.
It is available at http://review.typo3.org/6990

#10 - 2012-06-28 23:34 - Gerrit Code Review

Patch set 4 for branch **master** has been pushed to the review server.
It is available at http://review.typo3.org/6990

#11 - 2013-04-05 20:48 - Gerrit Code Review

Patch set 5 for branch **master** has been pushed to the review server.
It is available at https://review.typo3.org/6990

#12 - 2013-07-11 13:03 - Gerrit Code Review

Patch set 6 for branch **master** has been pushed to the review server.
It is available at https://review.typo3.org/6990

#13 - 2013-07-15 23:33 - Gerrit Code Review

Patch set 7 for branch **master** has been pushed to the review server.
It is available at https://review.typo3.org/6990

#14 - 2013-07-16 12:03 - Gerrit Code Review

Patch set 8 for branch **master** has been pushed to the review server.
It is available at https://review.typo3.org/6990

#15 - 2013-07-16 12:38 - Gerrit Code Review

Patch set 9 for branch **master** has been pushed to the review server.
It is available at https://review.typo3.org/6990

#16 - 2013-07-16 14:31 - Anonymous

- Status changed from Under Review to Resolved

- % Done changed from 0 to 100

Applied in changeset 819b5be0ac81004371fee2f0e6386cc32233a59b.

#17 - 2013-07-25 10:31 - Jigal van Hemert

- Tracker changed from Bug to Feature

- Subject changed from Links on external pages don't get indexed to Add possibility to start indexing an external site at a specific page

- Status changed from Resolved to New

- Assignee deleted (Dmitry Dulepov)

- Target version deleted (0)

- Complexity set to easy

- TYPO3 Version set to 4.0

#18 - 2013-07-29 09:59 - Michael Stucki

Hey Jigal,

why is this changed to new again? Did you revert the patch? Please explain...

Greetings, Michael

#19 - 2013-07-29 10:22 - Mario Rimann

Hi Michael

Jigal, Stefan Neufeind and I have discussed this issue during the last week and came to the conclusion, that my proposed patch would (in rare cases) break the existing functionality. We then discussed several ways of going forward:
a) revert my change + just leave as it was so far (= wait until someone really requires a))
b) revert my change + come up with a new proposal (then marked as feature as it would need to extend the crawler/indexed_search extensions in a way that won't fit as a bugfix)
c) revert my change + modify the patch so it would go through as bugfix (would lead to a "known unstable"-solution, which would probably fix 99.5% of all cases)

We threw away c) as it would not be "clean" at all. After some discussion, we decided to go for a) and just revoke + wait. And so did Jigal revert the change + update this issue.

I'd even go for closing this issue as "won't fix" for the moment - so that the bug-tracker get's cleaned up right away. If one really needs this change, he/she shall open a new report, referring to this one and we can get working on a proper solution).

Cheers,
Mario

#20 - 2013-07-29 10:29 - Michael Stucki

Thanks Mario! Now I see this was reverted in 559eb0091a5cf093515ad43d1b6b7dc7575bf8aa

Thanks for summarizing what happened about the issue. It's amazing to see how much work can go into such a tiny change... :-)

#21 - 2013-11-18 12:30 - Michael Stucki

- Status changed from New to Closed

Closed at request of Mario.

Files

indexed_search_4167_v1.diff	1.5 KB	2006-09-11	Administrator Admin
indexed_search_4167_v2.diff	1.4 KB	2009-08-28	Administrator Admin