TYPO3 Core - Task #89287

Make linkvalidator crawling polite

2019-09-26 12:40 - Sybille Peters

Status:	New	**Start date:**	2019-09-26
Priority:	Should have	**Due date:**	
Assignee:		**% Done:**	0%
Category:	Linkvalidator	**Estimated time:**	0.00 hour
Target version:			
TYPO3 Version:	10	**Complexity:**	
PHP Version:		**Sprint Focus:**	
Tags:	throttle, outgoing HTTP requests, resources, large-site		

Description

Currently, linkvalidator does not apply common practice for being "nice" / "polite" when crawling other websites:

- It should be possible to see what is crawling your site. It is usually standard to add a URL and contact information to the User-Agent or referrer, e.g.

```
"Mozilla/5.0 (compatible; MetaJobBot; http://www.metajob.de/crawler)"
```

```
"https://www.google.de/" "Mozilla/5.0 (iPhone; CPU iPhone OS 12_4_1 like Mac OS X) AppleWebKit/605
.1.15 (KHTML, like Gecko) Version/12.1.2 Mobile/15E148 Safari/604.1"
```

- the crawler should respect the robots.txt
- the crawler should wait between consecutive requests (Crawl-Delay). A good minimum value is e.g. 1-5 second for requests done on same domain

While linkvalidator is not a full web crawler which downloads the entire pages on the site and it currently uses HEAD by default (and not GET) which does not download the entire page - so this is not as dramatic.

But preferably, I think we should follow these recommendations as well.

Possible solutions

Not just for reducing load on other sites but also on the TYPO3 site, some changes might be made to the crawling process:

- when crawling external URLs, do not check right away but defer to a separate task which will handle the crawling of external URLs (with reasonable delays in between)
- do not keep crawling over and over but regularly only recrawl content which was recently modified
- optional: delegate the checking of external URLs, e.g. use a URLchecking service

This may make it necessary to make more things asynchronous and confine the link checking only to the scheduler.

Resources

About politeness of web crawlers:

- https://blog.scrapinghub.com/2016/08/25/how-to-crawl-the-web-politely-with-scrapy

URL checking site:

- https://httpstatus.io/

Related issues:

Related to TYPO3 Core - Epic #93547: Collection of problems with large sites	**Accepted**	**2021-02-19**

Associated revisions

Revision 2507a32f - 2019-11-17 00:11 - Sybille Peters

[FEATURE] Add additional configuration for external URLs

Additional configuration is added to customize settings for checking
external URLs.

Resolves: #86918
Related: #89287
Releases: master
Change-Id: I1ebfb31fe7760ad5b7c99db3999794c1e363cd17
Reviewed-on: https://review.typo3.org/c/Packages/TYPO3.CMS/+/61801
Tested-by: TYPO3com <noreply@typo3.com>
Tested-by: Chris Müller <typo3@krue.ml>
Tested-by: Benni Mack <benni@typo3.org>
Reviewed-by: Chris Müller <typo3@krue.ml>
Reviewed-by: Benni Mack <benni@typo3.org>

History

#1 - 2021-10-20 20:11 - Sybille Peters

- Tracker changed from Bug to Task

#2 - 2021-11-21 08:23 - Sybille Peters

- Tags set to throttle, outgoing HTTP requests, resources, large-site

#3 - 2021-11-21 08:23 - Sybille Peters

- Related to Epic #93547: Collection of problems with large sites added

#4 - 2021-11-21 08:30 - Sybille Peters

Not having a throttling of outgoing ~~URLs~~ HTTP requests and only caching the outgoing requests once per check cycle is one of the main reasons I do not want to use Linkvalidator in my site and I find that problematic if a TYPO3 extension bombards external sites with requests.

The way I currently changed the external link checking behaviour in "brofix" (which is a fork of linkvalidator with some changes), is that I added a minimum delay per domain: If the last request to an URL of this domain is less than X seconds, we wait. This delays the checking, but since I also use a link target cache and don't care if initial check takes several hours, I do see this as acceptable and an improvement in any case.

The wordpress plugin "Link checker" also has a throttle, see

- Disable domain based rate limiting
- https://github.com/wpmudev/broken-link-checker/blob/eedc557edcabe8893c9a7f15ccbcc72e0013fd72/modules/checkers/http.php#L31