EXT: Crawler

Extension Key: crawler

Language: en

Keywords: forAdmins, forDevelopers

The content of this document is related to TYPO3

- a GNU/GPL CMS/Framework available from www.typo3.org

Table of Contents

Introduction

What does it do?

Uses a command line cron-script to traverse a queue of actions (alternative option: executing all immediately), eg. requesting a URL but could also be calling of a function. The name “crawler” refers to this action: That a URL in the queue is requested.

Features an API that other extensions can plug into. Example of this is “indexed_search” which uses crawler to index content defined by its Indexing Configurations. Other extensions supporting it are “staticpub” (publishing to static pages) or “cachemgm” (allows recaching of pages).

The requests of URLs is specially designed to request TYPO3 frontends with special processing instructions. The requests sends a TYPO3 specific header in the GET requests which identifies a special action. For instance the action requested could be to publish the URL to a static file or it could be to index its content - or re-cache the page. These processing instructions are also defined by third-party extensions (and indexed search is one of them). In this way a processing instruction can instruct the frontend to perform an action (like indexing, publishing etc.) which cannot be done with a request from outside.

What's new?

Multiprocess support was added to the crawler to speed up the processing of queue entries.

Crawling of **workspaces** is now possible.

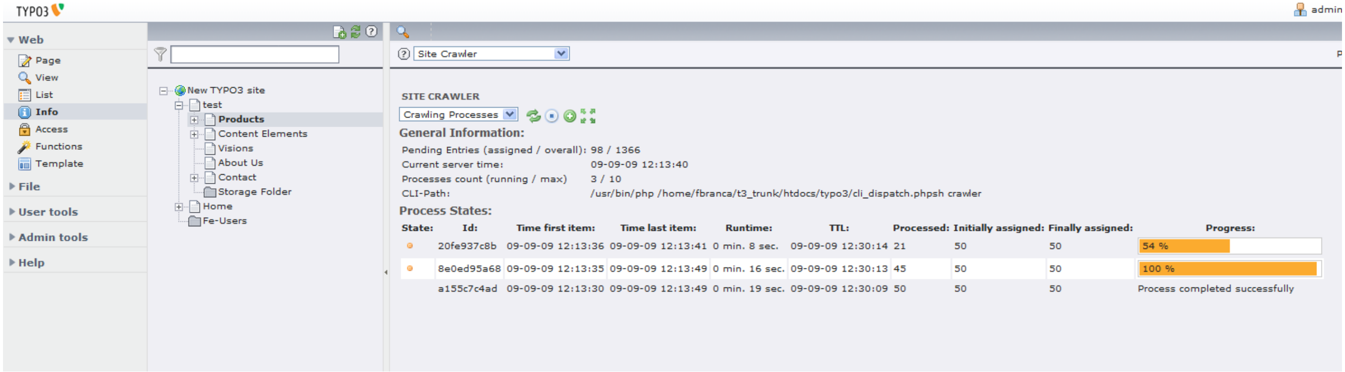
Some parts were rewritten using **MVC** principles.

Intotroduced “**pollable processing instructions**” to enable tracking failures in processing instructions. (Read more about this feature in chapter “pollable processing instructions”)

Screenshots

Has a backend module which displays the queue and log and allows execution and status check of the “cronscript” from the backend for testing purposes.

Here is the CLI (Command Line Interface = shell script = cron script) status display:



SITE CRAWLER

General Information:

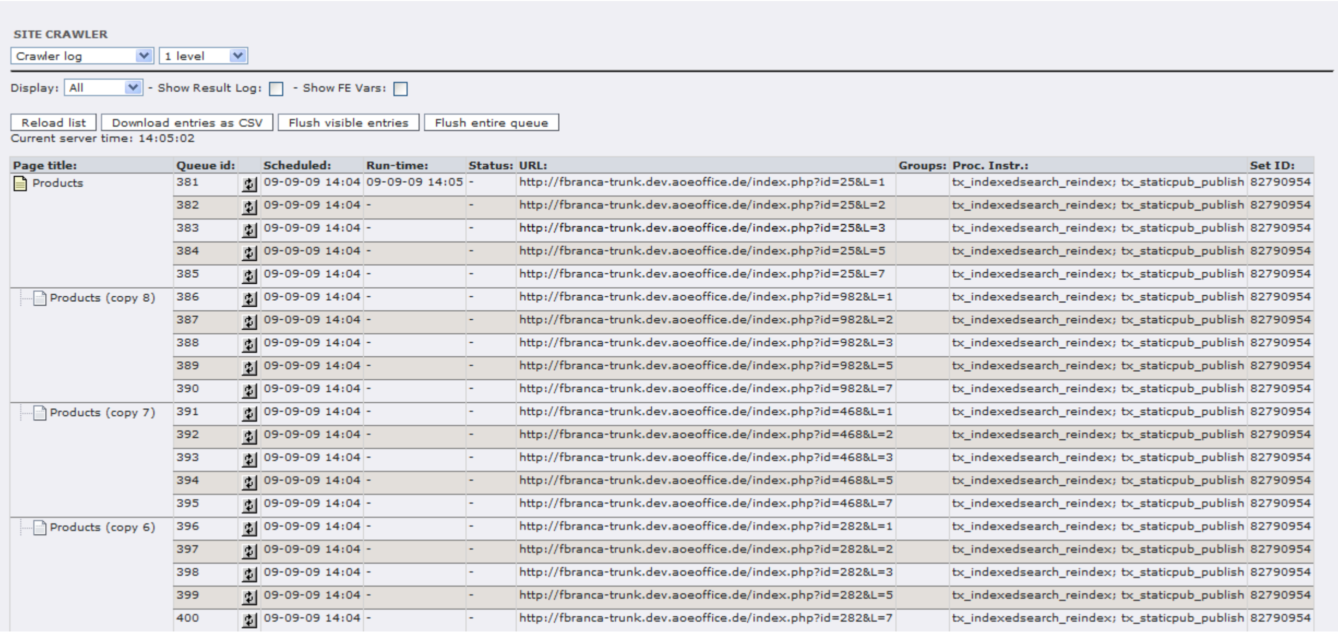
```
Pending Entries (assigned / overall): 98 / 1366
Current server time:                  09-09-09 12:13:40
Processes count (running / max)       3 / 10
CLI-Path:                             /usr/bin/php /home/fbranca/t3_trunk/htdocs/typo3/cli_dispatch.phpsh crawler
```

Process States:

State:	Id:	Time first item:	Time last item:	Runtime:	TTL:	Processed:	Initially assigned:	Finally assigned:	Progress:
●	20fe937c8b	09-09-09 12:13:36	09-09-09 12:13:41	0 min. 8 sec.	09-09-09 12:30:14	21	50	50	54 %
●	8e0ed95a68	09-09-09 12:13:35	09-09-09 12:13:49	0 min. 16 sec.	09-09-09 12:30:13	45	50	50	100 %
	a155c7c4ad	09-09-09 12:13:30	09-09-09 12:13:49	0 min. 19 sec.	09-09-09 12:30:09	50	50	50	Process completed successfully

Here is the crawler queue (before processing) / log (after processing)

Here is the interface for submitting a batch of URLs to be crawled. The parameter combinations are programmeble through Page Tsconfig or configuration records.

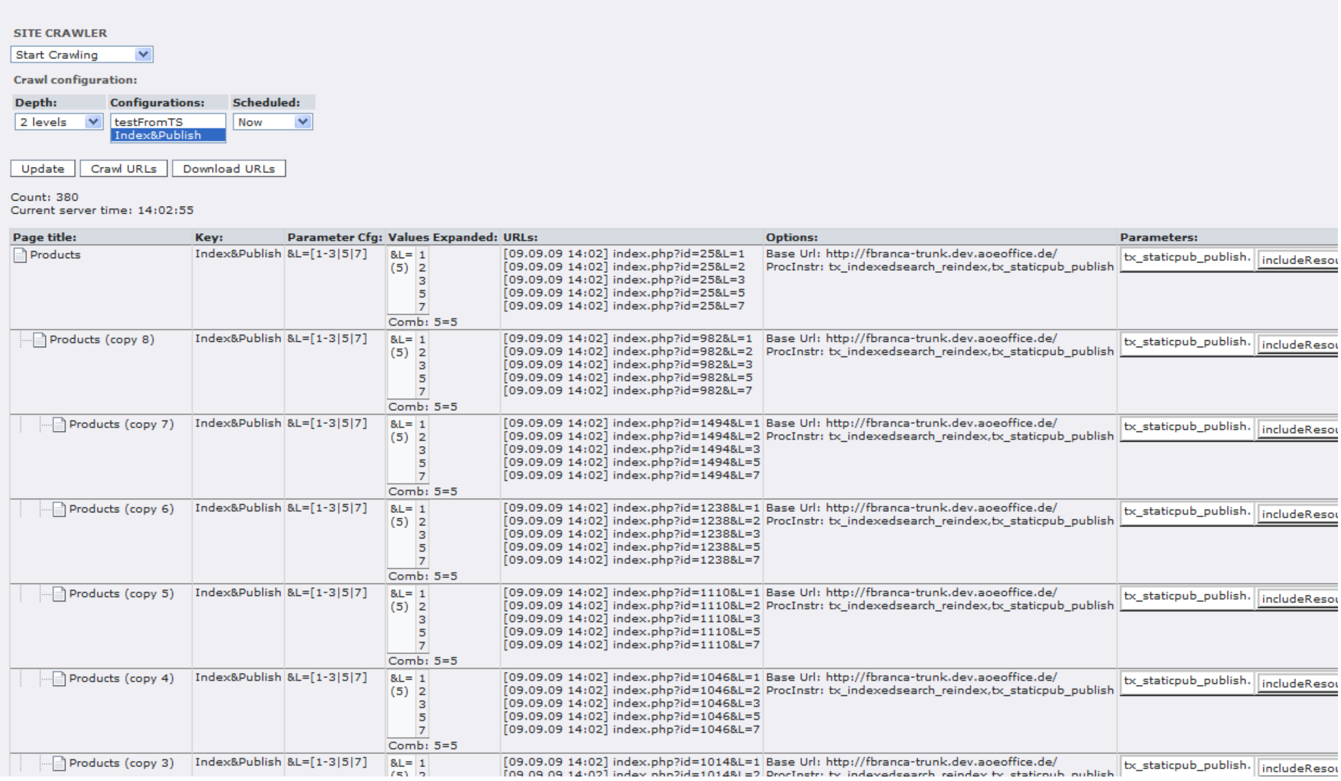


SITE CRAWLER

Display: All - Show Result Log: □ - Show FE Vars: □

Current server time: 14:05:02

Page title:	Queue id:		Scheduled:	Run-time:	Status:	URL:	Groups:	Proc. Instr.:	Set ID:
Products	381	↻	09-09-09 14:04	09-09-09 14:05	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=25&L=1		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	382	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=25&L=2		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	383	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=25&L=3		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	384	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=25&L=5		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	385	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=25&L=7		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
Products (copy 8)	386	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=982&L=1		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	387	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=982&L=2		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	388	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=982&L=3		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	389	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=982&L=5		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	390	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=982&L=7		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
Products (copy 7)	391	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=468&L=1		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	392	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=468&L=2		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	393	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=468&L=3		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	394	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=468&L=5		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	395	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=468&L=7		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
Products (copy 6)	396	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=282&L=1		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	397	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=282&L=2		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	398	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=282&L=3		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	399	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=282&L=5		tx_indexedsearch_reindex; tx_staticpub_publish	82790954
	400	↻	09-09-09 14:04	-	-	http://fbranca-trunk.dev.aoeoffice.de/index.php?id=282&L=7		tx_indexedsearch_reindex; tx_staticpub_publish	82790954



SITE CRAWLER

Crawl configuration:

Depth:	Configurations:	Scheduled:
2 levels	testFromTS Index&Publish	Now

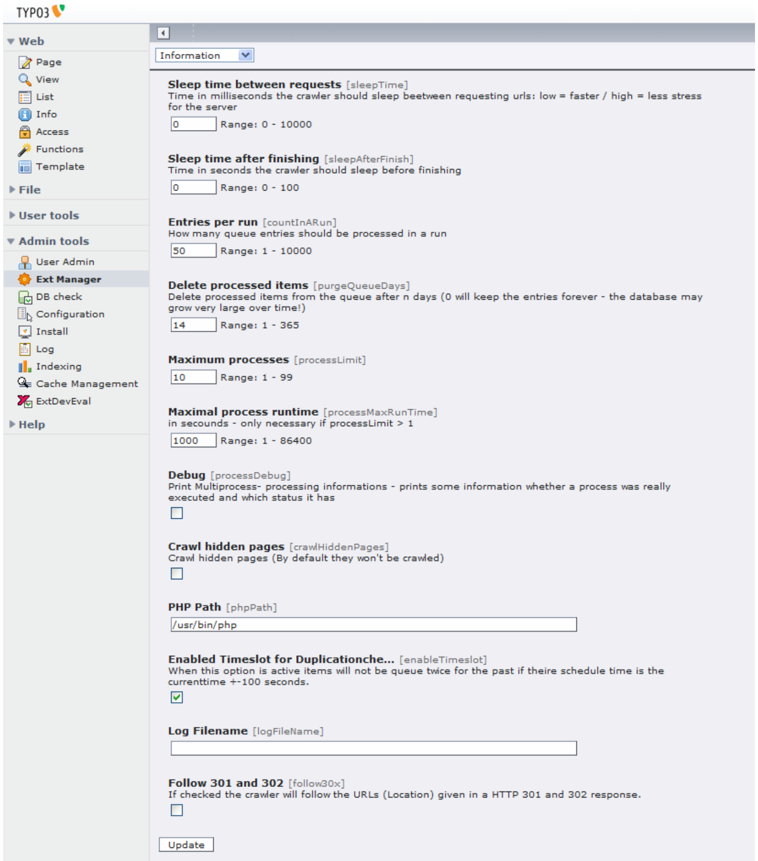
Count: 380
Current server time: 14:02:55

Page title:	Key:	Parameter Cfg:	Values Expanded:	URLs:	Options:	Parameters:
Products	Index&Publish	&L=[1-3|5|7]	&L= (5) 1 2 3 5 7 Comb: 5=5	[09.09.09 14:02] index.php?id=25&L=1 [09.09.09 14:02] index.php?id=25&L=2 [09.09.09 14:02] index.php?id=25&L=3 [09.09.09 14:02] index.php?id=25&L=5 [09.09.09 14:02] index.php?id=25&L=7	Base Url: http://fbranca-trunk.dev.aoeoffice.de/ ProcInstr: tx_indexedsearch_reindex,tx_staticpub_publish	tx_staticpub_publish. includeResou
Products (copy 8)	Index&Publish	&L=[1-3|5|7]	&L= (5) 1 2 3 5 7 Comb: 5=5	[09.09.09 14:02] index.php?id=982&L=1 [09.09.09 14:02] index.php?id=982&L=2 [09.09.09 14:02] index.php?id=982&L=3 [09.09.09 14:02] index.php?id=982&L=5 [09.09.09 14:02] index.php?id=982&L=7	Base Url: http://fbranca-trunk.dev.aoeoffice.de/ ProcInstr: tx_indexedsearch_reindex,tx_staticpub_publish	tx_staticpub_publish. includeResou
Products (copy 7)	Index&Publish	&L=[1-3|5|7]	&L= (5) 1 2 3 5 7 Comb: 5=5	[09.09.09 14:02] index.php?id=1494&L=1 [09.09.09 14:02] index.php?id=1494&L=2 [09.09.09 14:02] index.php?id=1494&L=3 [09.09.09 14:02] index.php?id=1494&L=5 [09.09.09 14:02] index.php?id=1494&L=7	Base Url: http://fbranca-trunk.dev.aoeoffice.de/ ProcInstr: tx_indexedsearch_reindex,tx_staticpub_publish	tx_staticpub_publish. includeResou
Products (copy 6)	Index&Publish	&L=[1-3|5|7]	&L= (5) 1 2 3 5 7 Comb: 5=5	[09.09.09 14:02] index.php?id=1238&L=1 [09.09.09 14:02] index.php?id=1238&L=2 [09.09.09 14:02] index.php?id=1238&L=3 [09.09.09 14:02] index.php?id=1238&L=5 [09.09.09 14:02] index.php?id=1238&L=7	Base Url: http://fbranca-trunk.dev.aoeoffice.de/ ProcInstr: tx_indexedsearch_reindex,tx_staticpub_publish	tx_staticpub_publish. includeResou
Products (copy 5)	Index&Publish	&L=[1-3|5|7]	&L= (5) 1 2 3 5 7 Comb: 5=5	[09.09.09 14:02] index.php?id=1110&L=1 [09.09.09 14:02] index.php?id=1110&L=2 [09.09.09 14:02] index.php?id=1110&L=3 [09.09.09 14:02] index.php?id=1110&L=5 [09.09.09 14:02] index.php?id=1110&L=7	Base Url: http://fbranca-trunk.dev.aoeoffice.de/ ProcInstr: tx_indexedsearch_reindex,tx_staticpub_publish	tx_staticpub_publish. includeResou
Products (copy 4)	Index&Publish	&L=[1-3|5|7]	&L= (5) 1 2 3 5 7 Comb: 5=5	[09.09.09 14:02] index.php?id=1046&L=1 [09.09.09 14:02] index.php?id=1046&L=2 [09.09.09 14:02] index.php?id=1046&L=3 [09.09.09 14:02] index.php?id=1046&L=5 [09.09.09 14:02] index.php?id=1046&L=7	Base Url: http://fbranca-trunk.dev.aoeoffice.de/ ProcInstr: tx_indexedsearch_reindex,tx_staticpub_publish	tx_staticpub_publish. includeResou
Products (copy 3)	Index&Publish	&L=[1-3|5|7]	&L= (5) 1 2	[09.09.09 14:02] index.php?id=1014&L=1 [09.09.09 14:02] index.php?id=1014&L=2	Base Url: http://fbranca-trunk.dev.aoeoffice.de/ ProcInstr: tx_indexedsearch_reindex,tx_staticpub_publish	tx_staticpub_publish. includeResou

Configuration

Extension Manager Configuration

A lot of options were added to the extension manager configuration, that allow settings to improve and enable new crawler features:



Sleep time between requests [sleepTime]
Time in milliseconds the crawler should sleep beetween requesting urls: low = faster / high = less stress for the server
0 Range: 0 - 10000

Sleep time after finishing [sleepAfterFinish]
Time in seconds the crawler should sleep before finishing
0 Range: 0 - 100

Entries per run [countInARun]
How many queue entries should be processed in a run
50 Range: 1 - 10000

Delete processed items [purgeQueueDays]
Delete processed items from the queue after n days (0 will keep the entries forever - the database may grow very large over time!)
14 Range: 1 - 365

Maximum processes [processLimit]
10 Range: 1 - 99

Maximal process runtime [processMaxRunTime]
in secounds - only necessary if processLimit > 1
1000 Range: 1 - 86400

Debug [processDebug]
Print Multiprocess- processing informations - prints some information whether a process was really executed and which status it has
□

Crawl hidden pages [crawlHiddenPages]
Crawl hidden pages (By default they won't be crawled)
□

PHP Path [phpPath]
/usr/bin/php

Enabled Timeslot for Duplicationche... [enableTimeslot]
When this option is active items will not be queue twice for the past if theire schedule time is the currenttime +-100 seconds.
☑

Log Filename [logFileName]

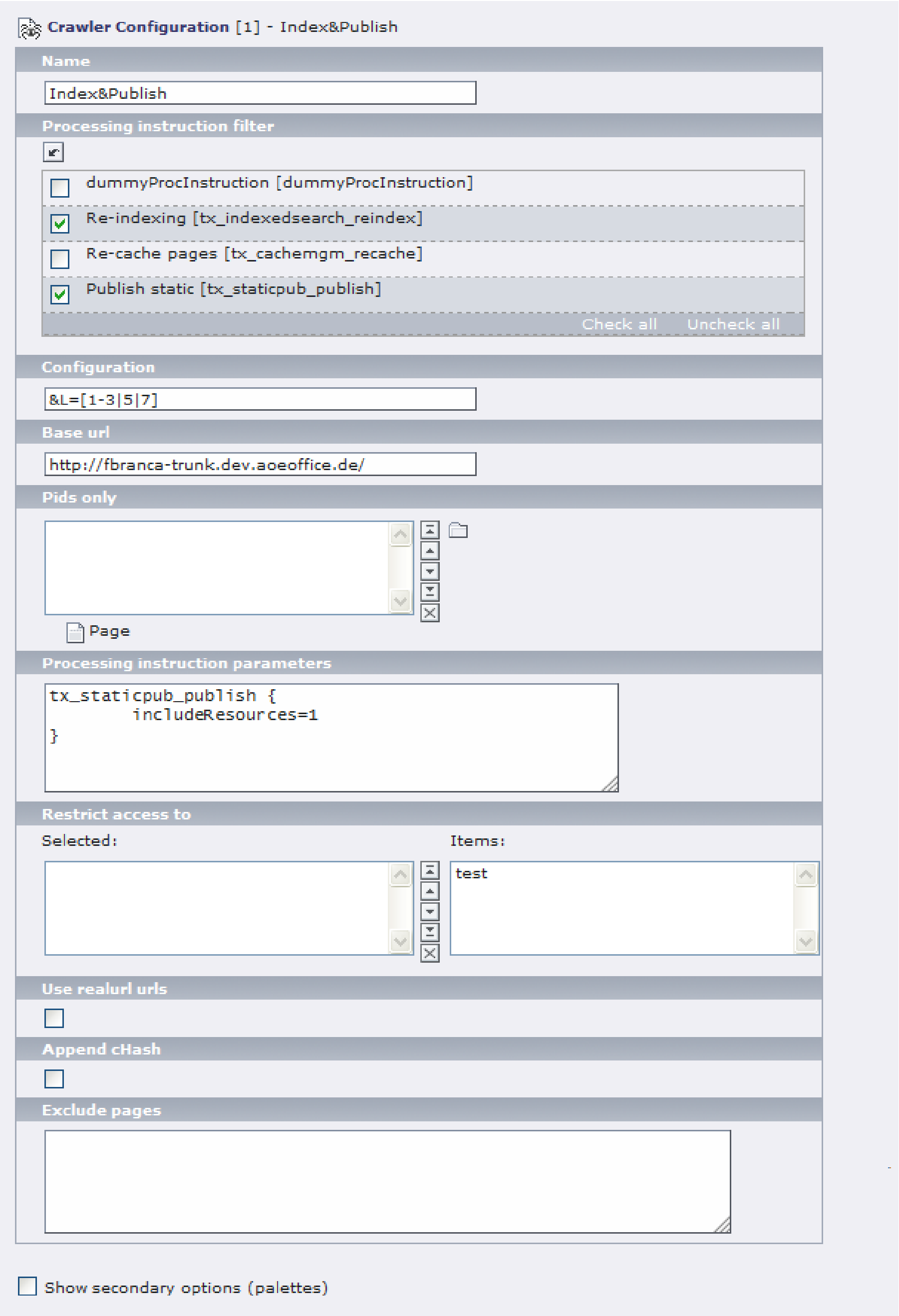
Follow 301 and 302 [follow30x]
If checked the crawler will follow the URLs (Location) given in a HTTP 301 and 302 response.
□

Update

Configuration records

Formerly configuration was done by using page ts (see below). This is still possible (fully backwards compatible) but not recommended. Instead of writing pagets simply create a configuration record (table: tx_crawler_configuration) and put it on the topmost page of the pagetree you want to affect with this configuration.

The fields in these records are related to the page ts keys described below. The "name" fields corresponds to the "key" in the pagets setup.



Crawler Configuration [1] - Index&Publish

Name

Index&Publish

Processing instruction filter

□ dummyProcInstruction [dummyProcInstruction]

☑ Re-indexing [tx_indexedsearch_reindex]

□ Re-cache pages [tx_cachemgm_recache]

☑ Publish static [tx_staticpub_publish]

Configuration

&L=[1-3|5|7]

Base url

http://fbranca-trunk.dev.aoeoffice.de/

Pids only

Processing instruction parameters

```
tx_staticpub_publish {
        includeResources=1
}
```

Restrict access to

Selected:

Items:

test

Use realurl urls

□

Append cHash

□

Exclude pages

□ Show secondary options (palettes)

Page TSconfig Reference (tx_crawler.crawlerCfg)

Property:	Data type:	Description:	Default:

paramSets.[key]	string	Get Parameter configuration. The values of GET variables are according to a special syntax. From the code documentation (class.tx_crawler_lib.php):
		• Basically: If the value is wrapped in [...] it will be expanded according to the following syntax, otherwise the value is taken literally
		• Configuration is splitted by "|" and the parts are processed individually and finally added together
		• For each configuration part:
		    • "[int]-[int]" = Integer range, will be expanded to all values in between, values included, starting from low to high (max. 1000). Example "1-34" or "-40--30"
		    • "**_TABLE**:” in the beginning of string indicates a look up in a table. Syntax is a string where [keyword]:[value] pairs are separated by semi-colon. Example "_TABLE:tt_content; _PID:123"
		        • Keyword “**_TABLE”** (mandatory, starting string): Value is table name from TCA to look up into.
		        • Keyword “**_PID**”: Value is optional page id to look in (default is current page).
		        • Keyword “**_FIELD**”: Value is field name to use for the value (default is uid).
		        • Keyword “**_PIDFIELD**”: Optional value that contains the name of the column containing the pid. By default this is “pid”.
		        • Keyword “**_ENABLELANG**”: Optional flag. If set only the records from the current language are fetched.
		    • - Default: Literal value
		
		Examples:
		`&L=[|1|2|3]`
		
		`&L=[0-3]`	

paramSets. [key].procInstrFilter	string	List of processing instructions, eg. “tx_indexedsearch_reindex” from indexed_searchto send for the request. Processing instructions are necessary for the request to perform any meaningful action, since they activate third party activity.	
paramSets. [key].procInstrParams. [procIn.key].[...]	strings	Options for processing instructions. Will be defined in the respective third party modules. **Examples:** .....procInstrParams.tx_staticpub_publish.includeResources=1	
paramSets. [key].pidsOnly	list of integers (pages uid)	List of Page Ids to limit this configuration to	
paramSets. [key].userGroups	list of integers (fe_groups uid)	User groups to set for the request.	
paramSets.[key].cHash	boolean	If set, a cHash value is calculated and added to the URLs.	
paramSets. [key].baseUrl	string	If not set, t3lib_div::getIndpEnv('TYPO3_SITE_URL') is used to request the page. MUST BE SET if run from CLI (since TYPO3_SITE_URL does not exist in that context!)	

[Page TSconfig: tx_crawler.crawlerCfg]

Example

```
tx_crawler.crawlerCfg.paramSets.test = &L=[0-3]
tx_crawler.crawlerCfg.paramSets.test {
	procInstrFilter = tx_indexedsearch_reindex
}
```

Executing the queue

The idea of the queue is that a large number of tasks can be submitted to the queue and performed over longer time. This could be interesting for several reasons;

- To spread server load over time.
- To time the requests for nightly processing
- And simply to avoid "max_execution_time" of PHP to limit processing to 30 seconds !

Executing queue with cron-job

A "cron-job" refers to a script that runs on the server with time intervals.

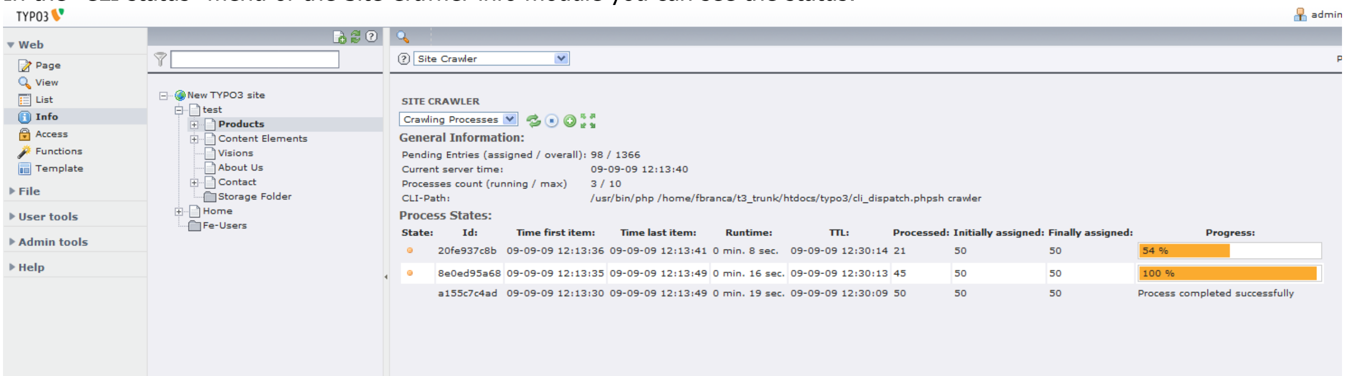
For this to become reality you must ideally have a cron-job set up. This assumes you are running on Unix architecture of some sort. The crontab is often edited by "crontab -e" and you should insert a line like this:

```
* * * * * [pathToYourTYPO3Installation]/typo3/cli_dispatch.phpsh crawler
```

This will run the script every minute. You should try to run the script on the command line first to make sure it runs without any errors. If it doesn't output anything it was successful.

You will have to add a user called "_cli_crawler" and you must have PHP installed as a CGI script as well in /usr/bin/

In the "CLI status" menu of the Site Crawler info module you can see the status:



SITE CRAWLER

Crawling Processes

General Information:

```
Pending Entries (assigned / overall): 98 / 1366
Current server time:                  09-09-09 12:13:40
Processes count (running / max)       3 / 10
CLI-Path:                             /usr/bin/php /home/fbranca/t3_trunk/htdocs/typo3/cli_dispatch.phpsh crawler
```

Process States:

State:	Id:	Time first item:	Time last item:	Runtime:	TTL:	Processed:	Initially assigned:	Finally assigned:	Progress:
●	20fe937c8b	09-09-09 12:13:36	09-09-09 12:13:41	0 min. 8 sec.	09-09-09 12:30:14	21	50	50	54 %
●	8e0ed95a68	09-09-09 12:13:35	09-09-09 12:13:49	0 min. 16 sec.	09-09-09 12:30:13	45	50	50	100 %
	a155c7c4ad	09-09-09 12:13:30	09-09-09 12:13:49	0 min. 19 sec.	09-09-09 12:30:09	50	50	50	Process completed successfully

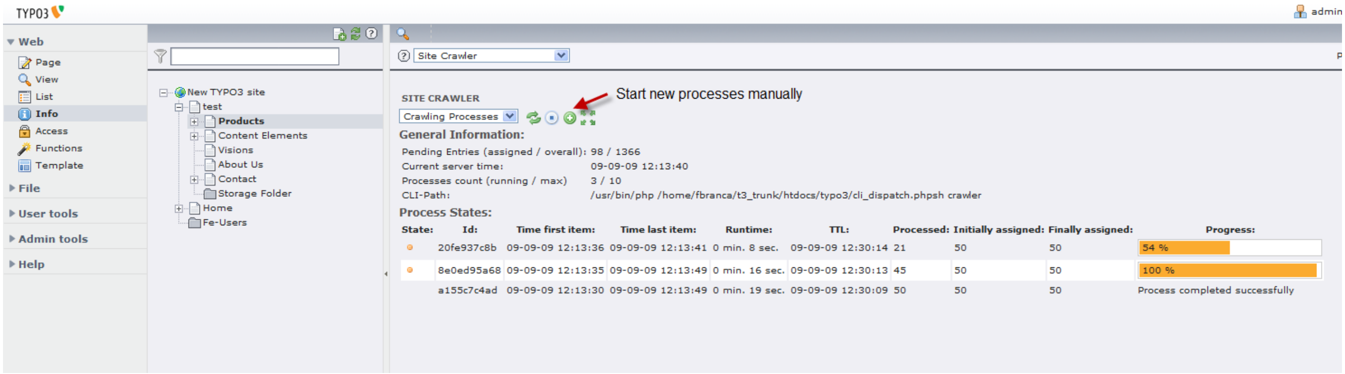
This is how it looks just after you ran the script. (You can also see the full path to the script in the bottom - this is the path to the script as you should use it on the command line / in the crontab)

If the cron-script stalls there is a default delay of 1 hour before a new process will announce the old one dead and run a new one. If a cron-script takes more than 1 minute and thereby overlaps the next process, the next process will NOT start if it sees that the "lock-file" exists (unless that hour has passed).

The reason why it works like this is to make sure that overlapping calls to the crawler CLI script will not run parallel processes. So the second call will just exit if it finds in the status file that the process is already running. But of course a crashed script will fail to set the status to "end" and hence this situation can occur.

Run via backend

To process the queue you must either set up a cron-job on your server or use the backend to execute the queue:



Start new processes manually

SITE CRAWLER

Crawling Processes

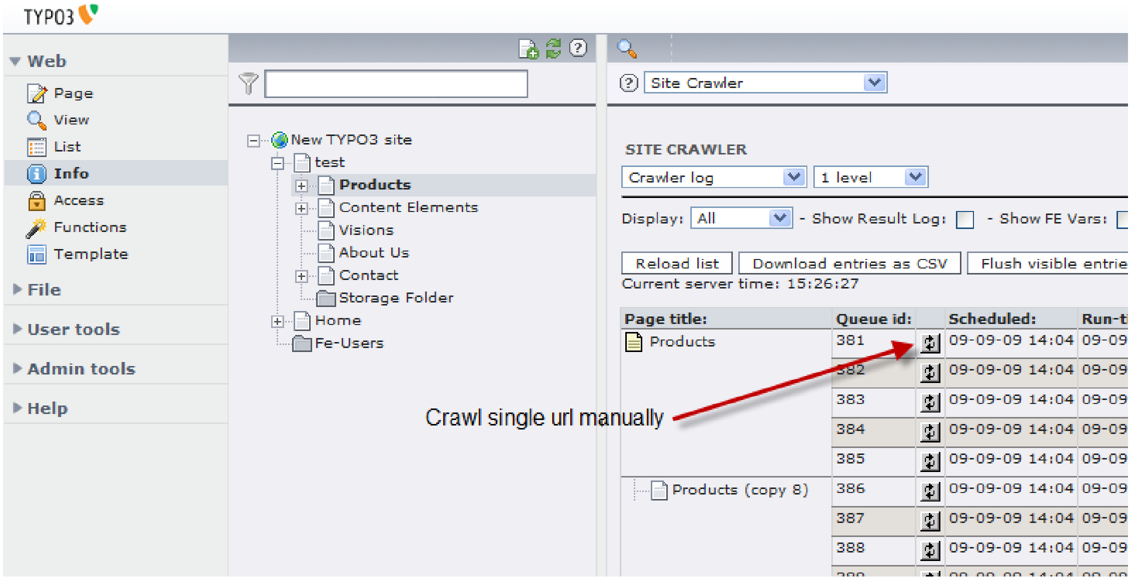
General Information:

```
Pending Entries (assigned / overall): 98 / 1366
Current server time:                  09-09-09 12:13:40
Processes count (running / max)       3 / 10
CLI-Path:                             /usr/bin/php /home/fbranca/t3_trunk/htdocs/typo3/cli_dispatch.phpsh crawler
```

Process States:

State:	Id:	Time first item:	Time last item:	Runtime:	TTL:	Processed:	Initially assigned:	Finally assigned:	Progress:
●	20fe937c8b	09-09-09 12:13:36	09-09-09 12:13:41	0 min. 8 sec.	09-09-09 12:30:14	21	50	50	54 %
●	8e0ed95a68	09-09-09 12:13:35	09-09-09 12:13:49	0 min. 16 sec.	09-09-09 12:30:13	45	50	50	100 %
	a155c7c4ad	09-09-09 12:13:30	09-09-09 12:13:49	0 min. 19 sec.	09-09-09 12:30:09	50	50	50	Process completed successfully

You can also (re-)crawl singly urls manually from within the Crawler log view in the info module:



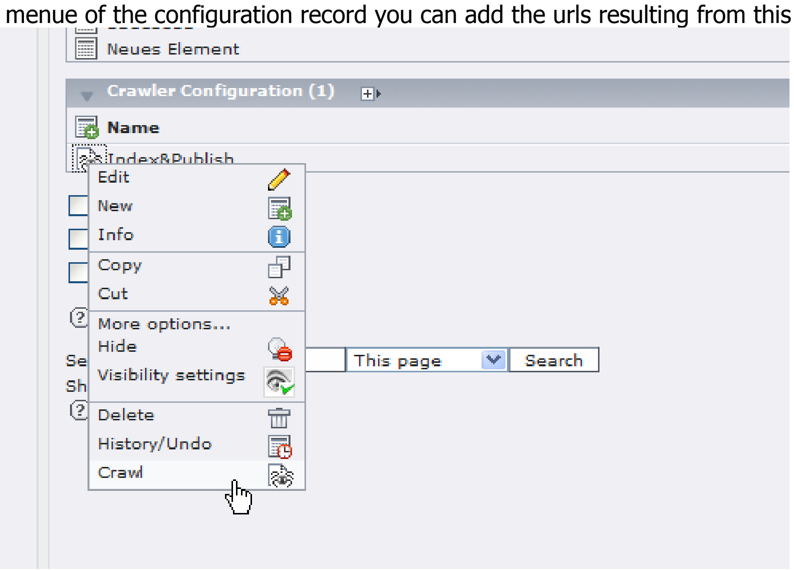
Current server time: 15:26:27

Page title:	Queue id:	Scheduled:	Run-t
Products	381	09-09-09 14:04	09-09
	382	09-09-09 14:04	09-09
	383	09-09-09 14:04	09-09
	384	09-09-09 14:04	09-09
	385	09-09-09 14:04	09-09
Products (copy 8)	386	09-09-09 14:04	09-09
	387	09-09-09 14:04	09-09
	388	09-09-09 14:04	09-09

Crawl single url manually

Adding entries to the queue by context menue

By clicking on the context menue of the configuration record you can add the urls resulting from this record to the queue:



Building and Executing queue right away (from cli)

An alternative mode is to automatically build and execute the queue from the command line in one process. This doesn't allow scheduling of task processing and consumes as much CPU as it can. On the other hand the job is done right away. In this case the queue is both built and executed right away.

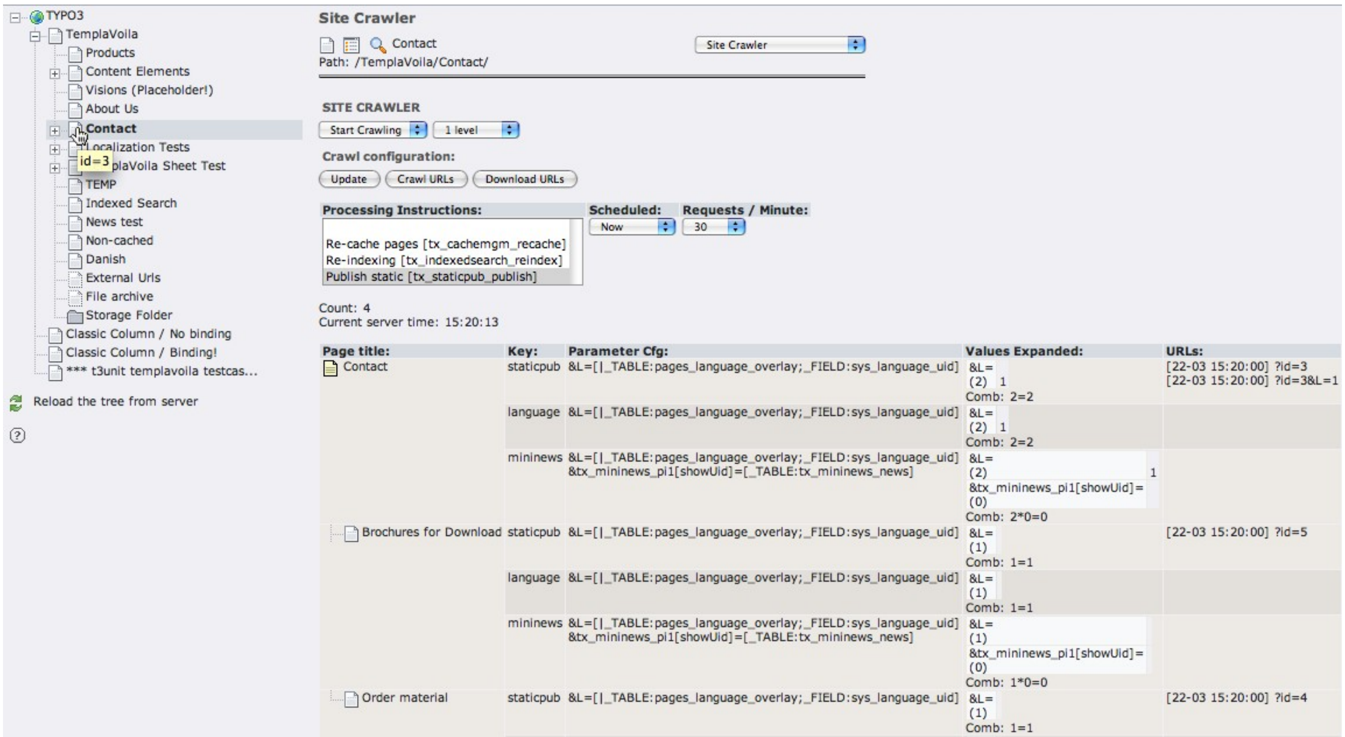
The script to use is this:

```
[pathToYourTYPO3Installation]/typo3/cli_dispatch.phpsh crawler_im
```

If you run it you will see a list of options which explains usage.

Property:	Data type:	Description:	Default:
-conf configurationKeys	string	Configurationkey: Commaseperated list of your crawler configurations. If you use the crwaler configuration records you have to use the "title" if your still using the old TypoScript based configuration you have to use the configuration key which is also a string. **Examples:** `-conf re-crawle-pages,re-crawle-news`	n/a
-n number	integer	Specifies how many items are put in the queue per minute. Only valid for output mode "queue".	n/a
-o mode	string	Output mode: "url", "exec", "queue" - url : Will list URLs which wget could use as input. - queue: Will put entries in queue table. - exec: Will execute all entries right away!	Queue
-d depth	integer	Tree depth, 0-99. How many levels under the 'page_id' to include.	n/a

Basically you must pass options similar to those you would otherwise select using the Site Crawler when you set up a crawler job ("Start Crawling"). Here is an example:



Site Crawler

Path: /TemplaVoila/Contact/

SITE CRAWLER

Start Crawling | 1 level

Crawl configuration:

Processing Instructions:

Re-cache pages [tx_cachemgm_recache]
Re-indexing [tx_indexedsearch_reindex]
Publish static [tx_staticpub_publish]

Scheduled: Now

Requests / Minute: 30

Count: 4
Current server time: 15:20:13

Page title:	Key:	Parameter Cfg:	Values Expanded:	URLs:
Contact	staticpub	&L=[|_TABLE:pages_language_overlay;_FIELD:sys_language_uid]	&L= (2) 1 Comb: 2=2	[22-03 15:20:00] ?id=3 [22-03 15:20:00] ?id=3&L=1
	language	&L=[|_TABLE:pages_language_overlay;_FIELD:sys_language_uid]	&L= (2) 1 Comb: 2=2	
	mininews	&L=[|_TABLE:pages_language_overlay;_FIELD:sys_language_uid] &tx_mininews_pi1[showUid]=[_TABLE:tx_mininews_news]	&L= (2) 1 &tx_mininews_pi1[showUid]= (0) Comb: 2*0=0	
Brochures for Download	staticpub	&L=[|_TABLE:pages_language_overlay;_FIELD:sys_language_uid]	&L= (1) Comb: 1=1	[22-03 15:20:00] ?id=5
	language	&L=[|_TABLE:pages_language_overlay;_FIELD:sys_language_uid]	&L= (1) Comb: 1=1	
	mininews	&L=[|_TABLE:pages_language_overlay;_FIELD:sys_language_uid] &tx_mininews_pi1[showUid]=[_TABLE:tx_mininews_news]	&L= (1) &tx_mininews_pi1[showUid]= (0) Comb: 1*0=0	
Order material	staticpub	&L=[|_TABLE:pages_language_overlay;_FIELD:sys_language_uid]	&L= (1) Comb: 1=1	[22-03 15:20:00] ?id=4

We want to publish pages under the page "ID=3 ("Contact" page selected) and 1 level down ("1 level" selected) to static files (Processing Instruction "Publish static [tx_staticpub_publish]" selected). Four URLs are generated based on the configuration (see right column in table).

To do the same with the CLI script you run this:

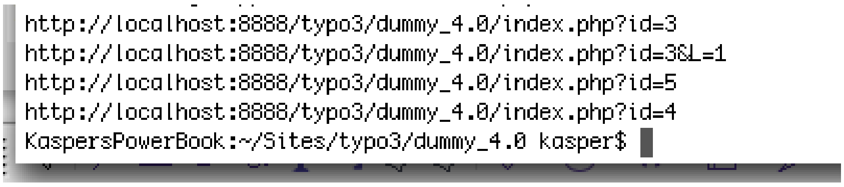
```
[pathToYourTYPO3Installation]/typo3/cli_dispatch.phpsh crawler_im 3 -d 1 -conf tx_staticpub_publish
```

And this is the output:

```
[22-03 15:29:00] ?id=3
[22-03 15:29:00] ?id=3&L=1
[22-03 15:29:00] ?id=5
[22-03 15:29:00] ?id=4
```

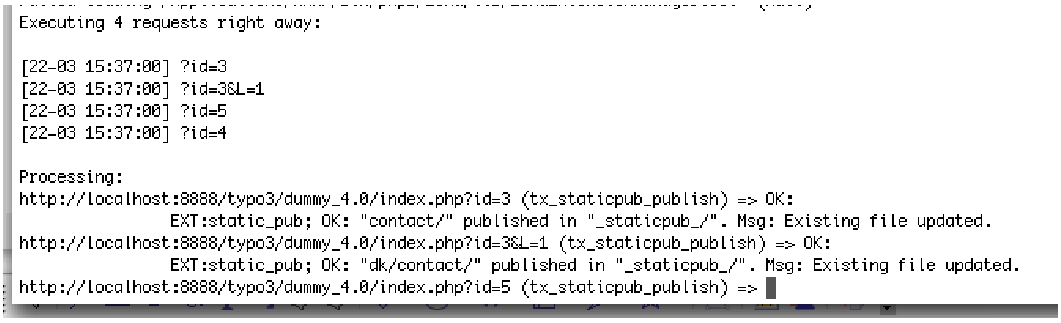
At this point you have three options for "action":

- Commit the URLs to the queue and let the cron script take care of it over time. In this case there is an option for setting the amount of tasks per minute if you wish to change it from the default 30. This is useful if you would like to submit a job to the cron script based crawler everyday.
  - Add "-o queue"
- List full URLs for use with wget or similar. Corresponds to pressing the "Download URLs" button in the backend module.
  - Add "-o url"



```
http://localhost:8888/typo3/dummy_4.0/index.php?id=3
http://localhost:8888/typo3/dummy_4.0/index.php?id=3&L=1
http://localhost:8888/typo3/dummy_4.0/index.php?id=5
http://localhost:8888/typo3/dummy_4.0/index.php?id=4
KaspersPowerBook:~/Sites/typo3/dummy_4.0 kasper$ █
```

- Commit and execute the queue right away. This will still put the jobs into the queue but execute them immediately. If server load is no issue to you and if you are in a hurry this is the way to go! It also feels much more like the "command-line-way" of things. And the status output is more immediate than in the queue.
  - Add "-o exec"



```
Executing 4 requests right away:

[22-03 15:37:00] ?id=3
[22-03 15:37:00] ?id=3&L=1
[22-03 15:37:00] ?id=5
[22-03 15:37:00] ?id=4

Processing:
http://localhost:8888/typo3/dummy_4.0/index.php?id=3 (tx_staticpub_publish) => OK:
                EXT:static_pub; OK: "contact/" published in "_staticpub_/". Msg: Existing file updated.
http://localhost:8888/typo3/dummy_4.0/index.php?id=3&L=1 (tx_staticpub_publish) => OK:
                EXT:static_pub; OK: "dk/contact/" published in "_staticpub_/". Msg: Existing file updated.
http://localhost:8888/typo3/dummy_4.0/index.php?id=5 (tx_staticpub_publish) => █
```

The examples above assume that "staticpub" is installed.

Features

Pollable processing instructions

Some processing instructions are never executed on the "client side" (the TYPO3 frontend that is called by the crawler). This happens for example if a try to staticpub a page containing non-cacheable elements. That bad thing about this is, that staticpub doesn't have any chance to tell that something went wrong and why. That's why we introduced the "pollable processing instructions" feature. You can define in the ext_localconf.php file of your extension that this extension should be "pollable" bye adding following line:

```
$GLOBALS['TYPO3_CONF_VARS']['EXTCONF']['crawler']['pollSuccess'][] = 'tx_staticpub';
```

In this case the crawler expects the extension to tell if everything was ok actively, assuming that something went wrong (and displaying this in the log) is no "success message" was found.

In your extension than simple write your "ok" status by calling this:

```
$GLOBALS['TSFE']->applicationData['tx_crawler']['success']['tx_staticpub'] = true;
```

Multiprocess support

If you want to optimize the crawling process for speed (instead of low server stress), maybe because the machine is a dedicated staging machine you should experiment with the new mulitprocess features.

In the exension manager you can set how many processes are allowed to run at the same time, how many queue entries a process should grab and how long a process is allowed to run. Then run one (or even more) crawling processes per minute. You'll be able to speed up the crawler by factor 10!

Hooks

excludeDoktype Hook

By adding doktype ids to following array you can exclude them from being crawled:

```
$GLOBALS['TYPO3_CONF_VARS']['EXTCONF']['crawler']['excludeDoktype'][] = <dokTypeId>
```

pageVeto Hook

You can also decide whether a page should not be crawled in an individual userfunction. Register your function here:

```
$GLOBALS['TYPO3_CONF_VARS']['EXTCONF']['crawler']['pageVeto'][] =
'EXT:yourext/.../class.tx_yourext_foo.php: tx_yourext_foo->bar';
```

FAQ

The crawler won't process all entrys at command-line-way. This might happened because the php run into an time out, to avoid this you can call the crawler like:

```
php -d max_execution_time=512 typo3/cli_dispatch.phpsh
```

The crawler won't process all entrys at command-line-way. This might happened because the php run into an time out, to avoid this you can call the crawler like: