



Recherche Indexée

Indexation automatique

# SOMMAIRE

<b>INTRODUCTION.....</b>	<b>3</b>
<b>1 L'EXTENSION INDEXED_SEARCH.....</b>	<b>3</b>
1.1 INSTALLATION .....	3
1.2 CONFIGURATION A L'INSTALLATION.....	3
1.3 CONFIGURATION DU TEMPLATE PRINCIPAL.....	3
1.4 CONFIGURATION DU TEMPLATE HTML.....	3
1.5 CREATION DE CONFIGURATION(S) D'INDEXATION .....	4
<b>2 L'EXTENSION CRAWLER.....</b>	<b>4</b>
2.1 INSTALLATION .....	4
2.2 CONFIGURATION .....	4
2.2.1 <i>Compte _CLI_crawler.....</i>	<i>4</i>
2.2.2 <i>TSconfig de la page racine du site .....</i>	<i>5</i>
2.2.3 <i>Définir les URLs à parcourir .....</i>	<i>5</i>
2.2.4 <i>Lancement du crawler.....</i>	<i>6</i>
<b>3 INDEXATION DE CONTENUS DE SITES DISTANTS .....</b>	<b>7</b>
<b>4 LE CRAWLER EN BACKEND.....</b>	<b>7</b>
<b>5 FORMULAIRE DE RECHERCHE .....</b>	<b>7</b>
5.1 SYSTEME PAR DEFAUT .....	7
5.2 UTILISATION DE L'EXTENSION MACINA_SEARCHBOX .....	8
5.3 AUTOCOMPLETION DES TERMES SAISIS .....	8
<b>6 PROBLEMES POSSIBLES .....</b>	<b>8</b>
6.1 PAS D'INDEXATION .....	8
6.2 FICHIERS A EXPLORER.....	8

# Introduction

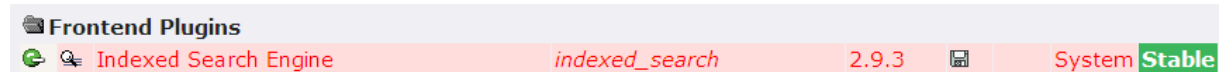
Ce document explique comment installer, configurer et utiliser la recherche indexée de Typo3.

Il explique comment utiliser l'extension *crawler* pour indexer les pages de façon automatique.

## 1 L'extension indexed\_search

### 1.1 Installation

Cette extension est intégrée à Typo3, il n'y a qu'à l'installer à partir du gestionnaire d'extensions :



### 1.2 Configuration à l'installation

- Full Text Data Length : si différent de 0, indique la taille limite des données indexées. Cela permet d'économiser de l'espace en base de données mais peut également limiter les possibilités de recherche.
- Disable Indexing in Frontend : par défaut, l'indexation des pages se fait en les consultant en frontend. Cocher cette option pour désactiver ce processus, dans le cas où on envisage une indexation lancée en backend ou par un CRON.

### 1.3 Configuration du template principal

Code à ajouter au setup du template de la page racine du site :

```
##### Recherche indexée #####
// L'indexation se fait sur les pages mises en cache => mettre les pages en cache
Page.config.no_cache = 0
// indexed search activée
page.config.index_enable = 1
// On utilise ce template
plugin.tx_indexedsearch.templateFile = fileadmin/templates/indexed_search.tmpl
// Ne pas afficher les explications
plugin.tx_indexedsearch.show.rules = 0
// Ne pas afficher le lien vers la recherche avancée
plugin.tx_indexedsearch.show.advancedSearchLink = 0
// Indexer des documents de type doc,pdf,...
// page.config.index_externals = 1
// ID de la page de départ de l'arborescence où on recherche (-1 => toute l'arborescence)
plugin.tx_indexedsearch.search.rootPidList = -1
// Afficher le num. des resultats
plugin.tx_indexedsearch.show.resultNumber = 1
// Si besoin de tt_news, autoriser leur mise en cache
plugin.tt_news.allowCaching = 1
```

### 1.4 Configuration du template HTML

Dans les templates HTML sur lesquels sont basées les pages du site, ajouter les marqueurs suivants autour des parties qui doivent être indexées :

```
<!--TYPO3SEARCH_begin-->
Mon Contenu
<!--TYPO3SEARCH_end-->
```

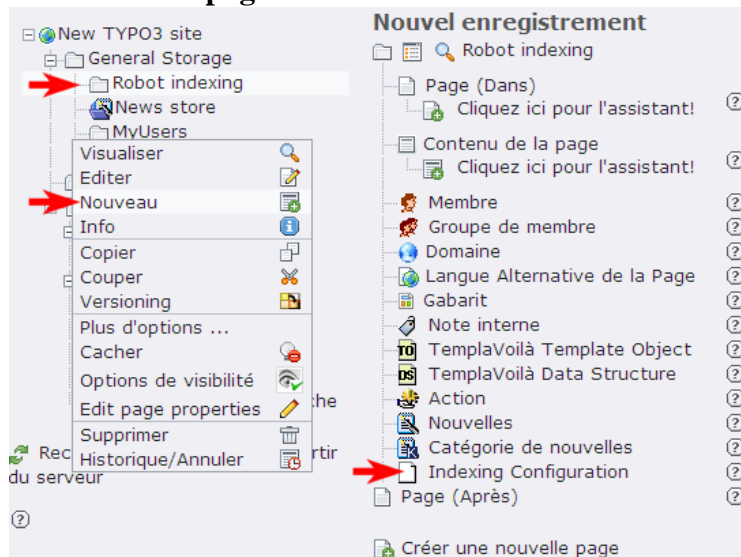
Si ces marqueurs sont absents, tout le contenu de la page sera indexé.

Il est possible d'utiliser plusieurs fois ce couple de marqueurs dans la même page, pour délimiter plusieurs zones à indexer.

## 1.5 Création de configuration(s) d'indexation

Lorsqu'on installe l'extension *indexed\_search*, il devient possible de créer des pages de type *indexing configuration*, qui stockeront des configurations d'indexation.

- Créer un dossier de type *sysfolder* qui stockera ces enregistrements
- Créer une page de type *indexing configuration*, qui stockera la configuration pour indexer les pages du site :



- Renseigner le titre, le type (pagetree), choisir la page racine (root page) et la profondeur à laquelle il faut étendre l'indexation, à partir de la page racine (depth).
- S'assurer que la case *disabled* est décochée (visible si on affiche les options secondaires)

**Note :** les enregistrements de ce type sont stockés dans la table *index\_config*.

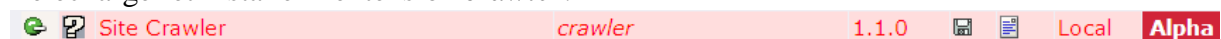
## 2 L'extension crawler

L'extension *crawler* permet de parcourir des URL en effectuant un traitement. Nous allons l'utiliser dans le cadre de la recherche indexée pour parcourir les URL dont nous voulons indexer le contenu.

On peut se passer de cette extension si on envisage d'indexer les pages en allant les consulter manuellement en frontend (cas d'un site avec peu de pages dont le contenu change peu souvent).

### 2.1 Installation

Télécharger et installer l'extension *crawler*.



### 2.2 Configuration

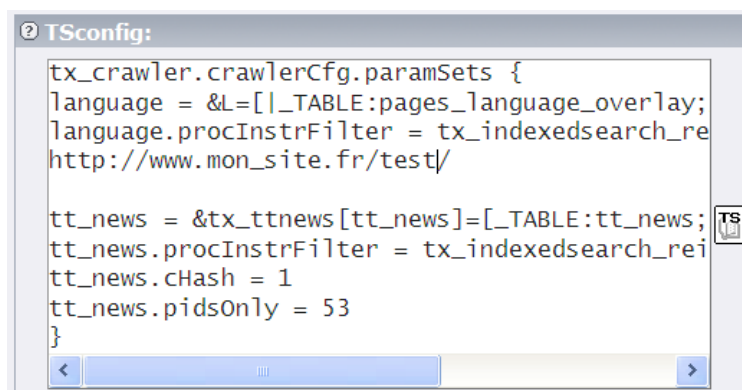
#### 2.2.1 Compte \_CLI\_crawler

Créer un compte utilisateur backend dont le login est **\_cli\_crawler**, le mot de passe peut être quelconque, ne renseigner aucun autre champ.

## 2.2.2 TSconfig de la page racine du site

A placer dans la partie *TSconfig* de la page racine du site, en adaptant au site concerné :

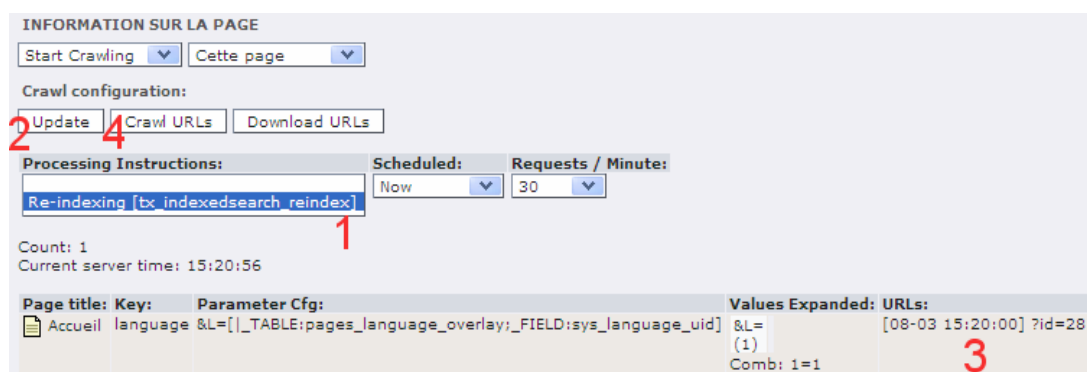
```
tx_crawler.crawlerCfg.paramSets {
language = &L=[_TABLE:pages_language_overlay;_FIELD:sys_language_uid]
language.procInstrFilter = tx_indexedsearch_reindex
// A adapter : URL du site
language.baseUrl = http://www.mon_site.fr/test/
// A adapter : _PID = ID du dossier contenant les news
tt_news = &tx_ttnews[tt_news]=[_TABLE:tt_news;_PID:51]
tt_news.procInstrFilter = tx_indexedsearch_reindex
tt_news.cHash = 1
// A adapter : ID de la page affichant une seule news
tt_news.pidsOnly = 53
}
```



## 2.2.3 Définir les URLs à parcourir

- Dans le module *Info*, choisir la page racine du site puis *site crawler* dans la liste déroulante du haut.
- Cliquer sur *Re-indexing [tx\_indexedsearch\_reindex]* pour indiquer le traitement à effectuer (1)
- Cliquer sur *update* (2), la partie URLs est renseignée (3)
- Cliquer sur *Crawl URLs* (4), s'assurer que le nombre d'URL correspond à notre attente et cliquer sur *continue*.

Cette opération alimente la table *tx\_crawler\_queue*, dans laquelle on peut éventuellement faire un suivi des opérations.



## 2.2.4 Lancement du crawler

Pour lancer le crawler (et donc l'indexation) en ligne de commande, ou par une tâche CRON :

- Modifier le fichier `typo3conf/ext/crawler/cli/crawler_cli.phpsh` en remplaçant la ligne : `define('PATH_thisScript',$_ENV['_']?$_ENV['_']:$_SERVER['_']);` par la ligne `define('PATH_thisScript',$_SERVER['argv'][0]);`
- Tester que le traitement fonctionne en lançant le crawler en ligne de commande, exemple sous windows :

```
C:\wamp\php>php.exe c:/wamp/www/t3_ia33/typo3conf/ext/crawler/cli/crawler_cli.phpsh
```

**Attention :** mentionner le chemin **absolu** du script `crawler_cli.phpsh`.

**Attention :** ce traitement peut nécessiter de modifier la directive de configuration de PHP `max_execution_time` (dans `php.ini`), pour que le script ait le temps de s'exécuter.

Lors de l'exécution du traitement, la planification de la prochaine indexation est faite (update de la table `tx_crawler_queue`).

On peut le voir en allant dans le module `liste/<dossier de stockage des configurations d'indexation>/<configuration d'indexation>`

On peut voir la date et l'heure à partir de laquelle la prochaine indexation pourra avoir lieu. Cela dépend de la valeur du facteur *how often would you like a re-index*.

Décocher la case *next indexing is scheduled* pour pouvoir relancer le crawler immédiatement.



On peut suivre le traitement en base dans la table `tx_crawler_queue` :

- 1 : la 1<sup>ère</sup> tâche prévue (enregistrée en base quand clic sur *crawl ULRs* en backend)
- 2 :
- 3 : le traitement crée un enregistrement qui planifie la prochaine tâche d'indexation, *scheduled* indique l'heure de l'exécution. Si on lance le crawler avant, le traitement ne se lancera pas.

Pour pouvoir relancer le traitement immédiatement, aller modifier la configuration d'indexation en backend.

parameters	scheduled	exec_time	set_id
1 x.fr/t3_ia33/index.php?id=28";s:16:"proclnstructions";a:1:{i:0;s:24:"tx_indexedsearch_reindex";}}	1173372066	1173430743	69611068
2 nstructions";a:1:{i:0;s:17:"[Index Cfg " _CALLBACKOBJ";s:62:"EXT:indexed_search/class.crawler.php&tx_indexedsearch_crawler";}	1173430743	1173430764	48234197
3 x.fr/t3_ia33/index.php?id=28";s:16:"proclnstructions";a:1:{i:0;s:24:"tx_indexedsearch_reindex";}}	1173430764	0	0

Exemple pour lancer le crawler avec un CRON : ajouter une des lignes suivantes à votre fichier de cron (édition : crontab -e)

```
# Toutes les 5 minutes, pour tests, impossible en prod
0,5,10,15,20,25,30,35,40,45,50,55 * * * * php
/web/Typo3/t3_ia33/typo3conf/ext/crawler/cli/crawler_cli.phpsh
```

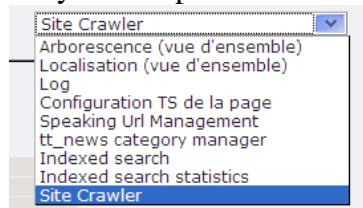
```
# Toutes les heures + 15 minutes
15 * * * * php /web/Typo3/t3_ia33/typo3conf/ext/crawler/cli/crawler_cli.phpsh
```

### 3 Indexation de contenus de sites distants

- Créer une nouvelle page de type *indexing configuration* dans le dossier sysfolder de stockage des configurations d'indexation
- Type : external URL
- External URL : donner l'URL de départ d'indexation, **en pensant à mentionner le slash final**, si besoin ([http://www.mon\\_site.fr/](http://www.mon_site.fr/))
- S'assurer que la case *disabled* est décochée (visible si on affiche les options secondaires)

### 4 Le crawler en backend

On y accède par le module web/info/<une\_page>/Site crawler



- Start crawling : configuration des pages à parcourir dans le site local
  - Faire un *update* (en cliquant au préalable sur *Re-indexing*) puis un *crawl URL* à chaque fois que la configuration TSconfig de la page est modifiée
- Crawler log : informations sur l'état d'indexation des URL à indexer
  - Flush entries : vide la table *tx\_crawler\_queue* et supprime donc les tâches d'indexation
- Cli status
  - status
    - Status = end : le traitement peut être lancé
    - Status = start : le traitement est en cours
  - On peut être dans un cas où le traitement a échoué et où le status vaut start. Dans ce cas, pour relancer le traitement, cliquer sur *disable* puis sur *enable*.
  - Cliquer sur *run now* pour lancer l'indexation depuis le backend

## 5 Formulaire de recherche

### 5.1 Système par défaut

L'utilisation la plus basique consiste à insérer le plugin *indexed\_search* comme élément dans une page, ce qui impose, si on utilise le plugin tel quel, de placer un lien dans un menu qui permet d'accéder à la page de recherche.

## 5.2 Utilisation de l'extension *macina\_searchbox*

Cette extension permet de placer facilement un formulaire de recherche sur toutes les pages du site.

Cette extension n'est pas un nouveau plugin de recherche, son fonctionnement est lié au plugin *indexed search box* qui doit donc être installé, configuré et inséré dans une page; quand on valide la recherche, la page appelée est celle qui contient ce plugin.

- Installer l'extension *macina\_searchbox\_2.2.0.t3x*
- Dans le module *template*, choisir *template analyser* (liste déroulante en haut à droite)
- Cliquer sur *macina\_searchbox*
- Copier la config de la partie *[global]*
- Passer en mode *info/modify*, coller le code dans le setup du template
- Renseigner *pidSearchpage* avec l'ID de la page qui contient le plugin *indexed search*
- Pour le code complet à intégrer, se baser sur le fichier *macina\_searchbox\_ts.txt*
- Pour modifier l'apparence du formulaire, modifier ce fichier :  
*typo3conf/ext/macina\_searchbox/pi1/template.htm*
- Pour modifier l'apparence du formulaire de *indexed search* et des résultats, modifier ce fichier : *typo3\sysext\indexed\_search\pi\indexed\_search.tpl*  
Le fichier de ressource *indexed\_search\_light.tpl* contient le code pour générer seulement l'affichage des résultats de la recherche, sans réafficher le formulaire de recherche.

## 5.3 Autocomplétion des termes saisis

Pour ajouter cette fonctionnalité :

- L'extension *indexed\_search* doit être installée
- Importer et installer l'extension *cb\_indexedsearch\_autocomplete\_0.3.0.t3x*
- Si l'extension *macina\_searchbox* est utilisée, modifier le template principal en ajoutant *indexed search autocomplete* à la partie *include static (from extensions)*.
- Eventuellement, modifier cette feuille de style pour personnaliser l'apparence de l'extension */res/cb\_indexedsearch\_autocomplete.css*

# 6 Problèmes possibles

## 6.1 Pas d'indexation

Si les pages ne s'indexent pas, essayer d'installer l'extension-patch suivante :

*dvdg\_indexedsearch\_patch\_0.0.1.t3x*

## 6.2 Fichiers à explorer

Les fichiers suivants contiennent des traitements relatifs au crawler et à la recherche indexée :

- *typo3conf/ext/crawler/class.tx\_crawler\_lib.php* : définition d'une classe de librairie de fonctions pour le crawler.
- *t3lib/class.t3lib\_div.php* : *fsockopen* sur des URL externes (ligne 2356), fonction *getUrl*