# How to…

# set up a Web Repository and Crawling it for Indexing

ENTERPRISE PORTAL 5.0 / 6.0

**PUBLIC**

---

---

## ASAP "How to…" Paper

**Applicable Releases: EP 5.0 SP5 / EP 6.0 SP0**

**April 2003**

1 SCENARIO ..... 1

2 THE STEP BY STEP SOLUTION ..... 1

2.1 Creating and Configuring a Web Repository Manager ..... 2

2.2 Configuring a Crawler Profile ..... 11

2.3 Creating an Index and Assigning a Web RM ..... 13

2.4 Monitoring Crawling and Indexing ..... 16

3 RESULT ..... 19

# 1 Scenario

You want to use SAP Enterprise Portal to search for documents that are stored on various Web sites. The Web server can be reachable in the Internet or in an Intranet.

The scenario assumes that you want to make the Internet pages of a news provider accessible through the portal. The Web server of the news provider is to be checked twice a day for new documents.

The documents can be indexed after the crawler has found them on the Web site. The documents are then available to portal users when they carry out a search. If you are using taxonomies, you can also place these documents into categories. Portal users then access the documents through a taxonomy.

# 2 The Step By Step Solution

You need a Web repository manager (Web RM) to access a Web server. The Web server acts as a data source. You can create a separate Web RM for each Web server, or you can group together several Web servers in a Web repository manager.

See Creating and Configuring a Web Repository Manager [Page 2].

Crawlers that search the Web servers for documents are used to provide the contents of linked Web servers in the portal.

In the KM platform, crawlers and crawler profiles are preconfigured for your use. You have to create a new crawler profile if you want to define how many document levels the crawler should pursue on the Web server.

See Configuring a Crawler Profile [Page 11].

You have to create an index for the Web repository before you can search for documents or classify them. In index administration, select the crawler profile and define how often the Web server is to be searched for new, updated, or deleted documents. After the documents have been crawled they are forwarded to the TREX search engine where they are indexed.

See Creating an Index and Assigning a Web RM [Page 13].

You can use the crawler monitor and the TREX monitor to monitor the status of crawling and indexing.

See Monitoring Crawling and Indexing [Page 16].

As soon as the indexing process has been completed, you can search for documents stored in the Web repository from the portal.

See Result [Page 19].

## Notes

If you are using EP 5.0, you find the KM configuration in the *KM Admin* workset.

In EP 6.0, you find the KM configuration under *System Administration* → *System Configuration* → *Knowledge Management* → *Configuration*.

Configuration parameters that are needed for a simple configuration are highlighted in the table. The other parameters are for enhanced configurations.

## 2.1 Creating and Configuring a Web Repository Manager

This section of the How To guide explains how to create a Web repository manager.

You want to be able to call up documents that are located under the Internet address **`www.cnn.com/TECH`** in the corresponding Web repository.

Carry out the following steps to create the Web repository manager and the corresponding components.

1. Register the Web server in the KM system landscape.
2. Create a Web site.
3. Optional: Define HTML property extractors.
4. Create a cache for the repository manager.
5. Create a Web repository manager and configure it.

### 1. Register the Web Server in the KM System Landscape

Before you create a Web site, you have to define the URL of the Web server in the form of an HTTP system in the KM system landscape.

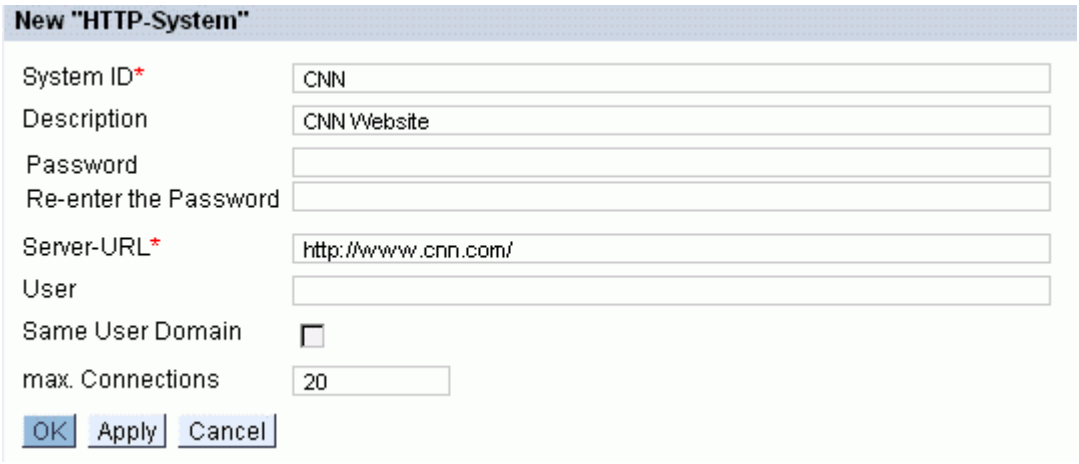
To create an HTTP system, choose *Content Management* → *Global Services* → *System Landscape Definitions* → *Systems* → *HTTP System*.

**Parameters of an HTTP System**

| Parameter | Required | Description |
|---|---|---|
| **`System ID`** | Yes | ID of the HTTP system.<br>Caution: Do not use spaces to separate elements of this ID. |
| `Description` | No | Description of the HTTP system. |
| `Password` | No | Specification of the password needed to access the HTTP system. |
| **`Server-URL`** | Yes | Specification of the URL that targets the HTTP system. |

| | | |
|---|---|---|
| `User` | No | Specification of the user needed to access the HTTP system. |
| `Same User Domain` | No | Specifies whether authentication information for SAP Enterprise Portal is forwarded to the remote server.<br>Activate this parameter if the HTTP system is to operate in the same user domain as SAP Enterprise Portal.<br>Deactivate this parameter if you want to access external systems (for example, an external Web site). |
| **`max. Connections`** | No | Specifies the maximum number of connections that a repository manager builds with this system.<br>The exact number of connections depends on the remote server. We recommend that you minimize the default settings. |

Example Configuration for an HTTP System




For more information on HTTP systems, see the KM administration guide.

### 2. Create a Web Site

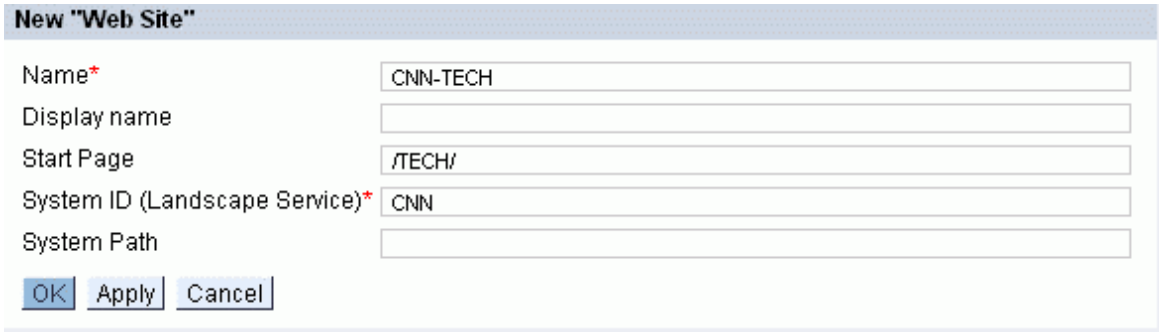
To create a Web site, choose *Content Management* → *Repository Managers* → *Web Sites.*

**Web Site Parameters**

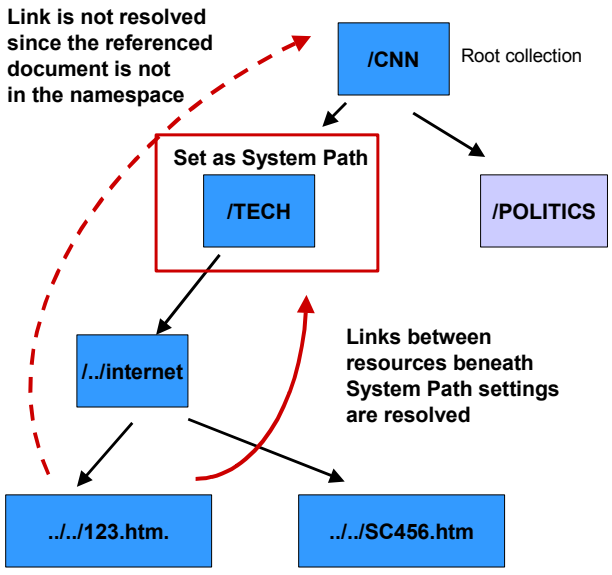
| Parameter | Required | Description |
|---|---|---|
| **`Name`** | Yes | Name of the Web site.<br>Caution: Do not use spaces to separate elements of this name. |
| `Display Name` | No | The display name for the Web site, which becomes the `displayname` property of the resource that is displayed in the browser. |
| **`Start Page`** | No | Explicitly specified start page that overrides the default start page of the Web site.<br>Enter the relative part of the URL for the server URL (see example). |

| | | |
|---|---|---|
| **`System ID (Landscape Service)`** | Yes | Specifies the *System ID* entered in the definition of the HTTP system. |
| `System Path` | No | Optional suffix for the server URL that is entered in the HTTP system.<br>This makes it possible to use the same system for different locations in the systems URL namespace. See the screenshot below. |

Example Configuration for a Web Site




Path Specifications for the Configuration of a Web Site.




If you use the parameter `Start Page` to define the initial point of entry to the repository, there is no limitation on the resolution of links.

If you use the parameter `System Path` (for example, /TECH), superordinate links at /CNN level are not pursued.

For more information on Web sites, see the KM administration guide.

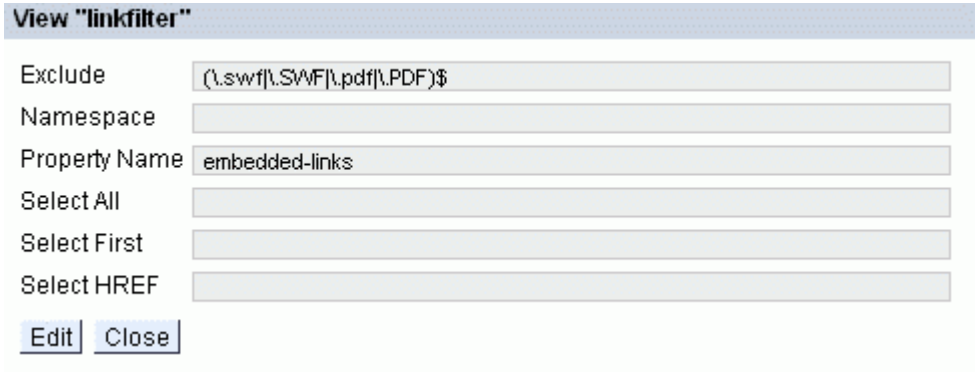
### 3. Optional: Define HTML Property Extractors

In our example, we want to define certain MIME types that a crawler should not pursue when analyzing the contents of a Web repository. We don't want Shockwave Flash and PDF files to be crawled. Normally, you would also want to exclude ZIP and MP3 files. You enter this information in the HTML property.

To create a new HTML property, choose *Content Management* → *Repository Managers* → *HTML Property Extractors* → *HTML Properties*.

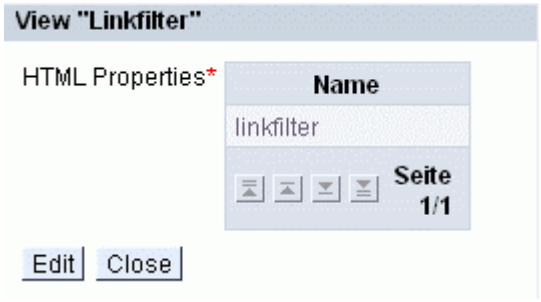
In the parameter `Exclude`, enter the MIME types to be excluded in the form of a *regular expression* (see screenshot). In the parameter `Property Name`, enter the value **`embedded-links.`**

Example Configuration for an HTML Property




You now select this HTML property in the definition of an HTML property extractor. To create a new HTML property, choose *Content Management* → *Repository Managers* → *HTML Property Extractors* → *HTML Property Extractor*.

Example Configuration for an HTML Property Extractor




You reference the HTML property extractor later on in the configuration for the Web repository manager.

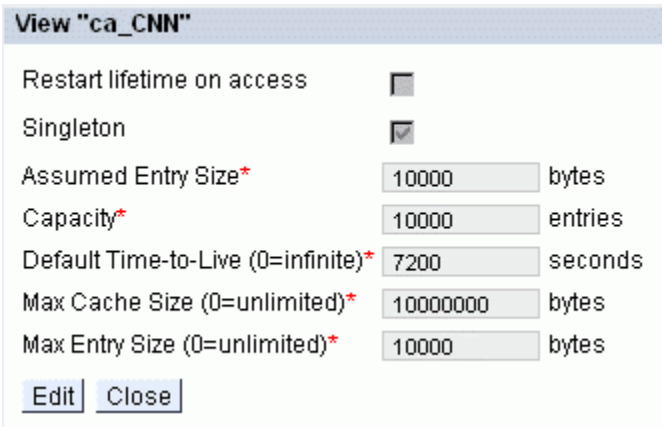
For more information on HTTP property extractors, see the KM administration guide.

### 4. Create a Cache for the Repository Manager

You create a separate cache for the Web repository manager.

To do this, choose *Content Management* → *Utilities* → *Caches* → *Memory Cache*.

Example Configuration for a Cache




For more information on caches, see the Cache How To and the KM administration guide.

### 5. Create a Web Repository Manager

To create a Web repository manager, choose *Content Management* → *Repository Managers* → *Web Repository*. In the configuration, you assign the Web site that you just created, the cache, and the HTML property extractor to the Web repository manager.

You have to enter a proxy server and port if

a) No access is possible to the Web site in the Intranet without them
b) You want to take advantage of the caching benefits of the proxy

You do not need to make entries for the parameters `Start Page`, `System ID (Landscape Service)` and `System Path`. `These parameters are only valid for` *`simple`* `Web repositories`. If you are using Web sites, you do not use these parameters.

Enter the following parameters in the configuration for a Web repository manager:

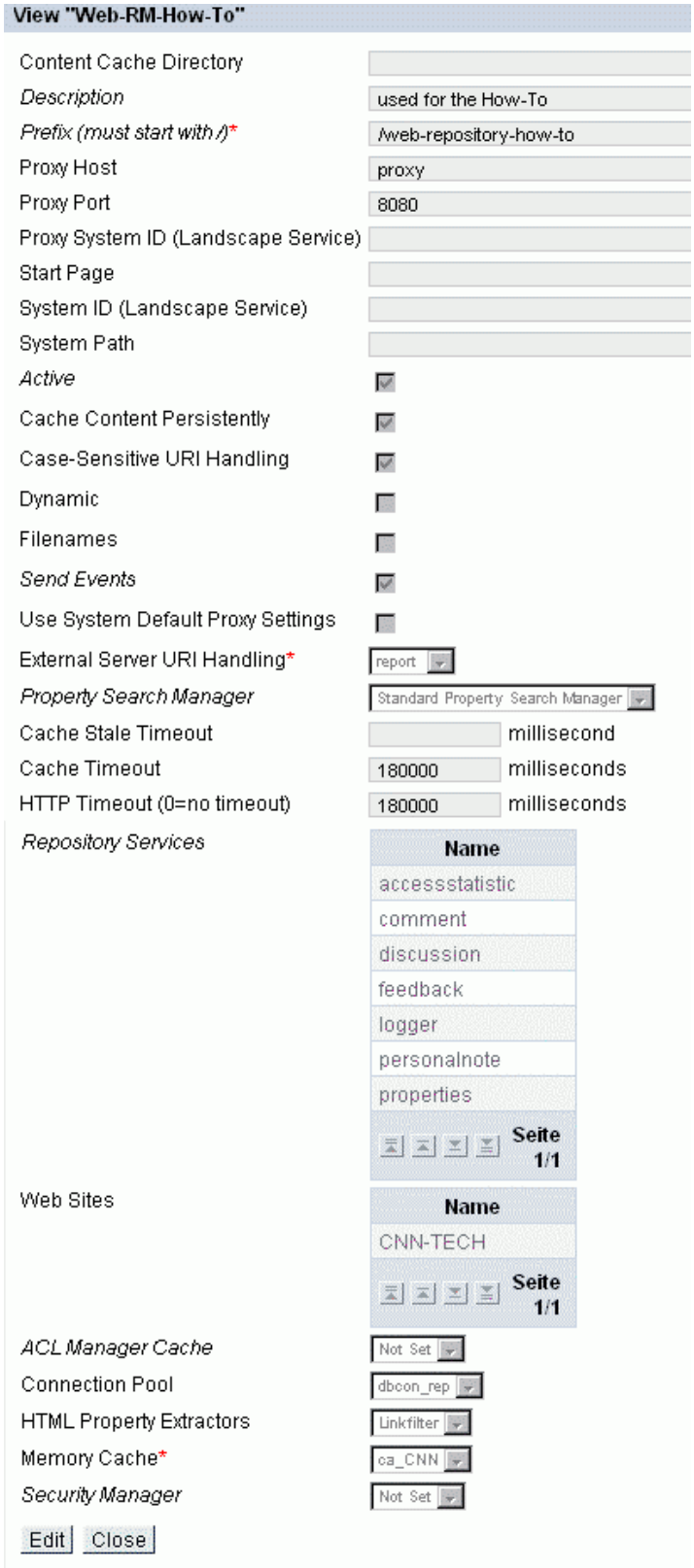
**Web Repository Manager Parameters**

| Parameter | Required | Description |
|---|---|---|
| **Name** | Yes | Name of the repository manager. |
| Content Cache Directory | No | Specifies a directory in the file system in which the resource contents for the cache are stored.<br>If no directory is specified, the resource contents are stored in the database. |
| Description | No | Description of the repository manager |
| **Prefix** | Yes | The URI prefix for which the manager is registered. This specification is entered in the list in the root directory.<br>Note that you must enter the prefix with a forward slash, for example, **/WEB REPOSITORY**. |
| **Proxy Host** | No | Hostname of an HTTP proxy to use for this repository<br>If the proxy server requires authentication, use the Proxy System ID parameter. |
| **Proxy Port** | No | Port number where the proxy is reachable. |
| Proxy System ID (Landscape Service) | No | Proxy server ID from the system landscape service<br>You should use this parameter instead of Proxy Host when the proxy server requires authentication. |
| Start Page | No | Explicitly specified start page that overrides the default start page of the Web site. |
| System ID (Landscape Service) | No | HTTP system specification that is registered in the CM system landscape.<br>You specify this parameter only if the repository is to map a **single Web site**. In this case, you must not select any of the Web sites that may be presented as options for the Web Sites property. |
| System Path | No | Optional suffix for the server URL of the remote system.<br>The server URL is a property of the (landscape) system. System_Path enables you to use the same system for different paths in the URL namespace. |
| Cache Content Persistently | No | Specifies whether resource content is cached.<br>Needs a parameter specification in the Connection Pool parameter.<br>Depending on the Content Cache Directory, the content is stored either in the database or in the file system. |

| | | |
|---|---|---|
| `Case-Sensitive URI Handling` | No | Determines whether the repository manager takes account of lowercase and uppercase text in resource URLs.<br>If this is active, the system distinguishes between *INDEX.HTM* and *index.htm*, for example. |
| `Dynamic` | No | Determines whether the repository can dynamically add new remote HTTP servers to its namespace. |
| `Filenames` | No | Determines whether the repository generates resource names that are valid file names.<br>Links inside an HTML page also use such names. The default value is `false`. |
| **Send Events** | No | Specifies whether the repository sends events when operations such as *delete* and *update content* are performed.<br>The repository sends events if this parameter is activated. |
| **Use System Default Proxy Settings** | No | Determines whether the Java Virtual Machine (VM) standard values `http.proxyHost` and `http.proxyPort` are taken into account.<br>You can set these values when starting the VM. If this parameter is activated, the settings apply to all Web repositories provided you have not configured any other settings for `Proxy Host` or `Proxy System ID`. |
| `External Server URI Handling` | Yes | Determines how URIs (links) inside HTML pages that point to external servers (outside the scope of this web repository) are handled.<br>*none*: URIs are transferred unchanged.<br>*rewrite*: URIs are rewritten to point to URIs on this server wherever possible.<br>*report*: Like *rewrite*, but the newly written URIs are also transmitted to the property `embedded-links` for the resource. |
| `Property Search Manager` | No | Selection of manager for property search.<br>It is used by services and applications that need to find resources based on their properties.<br>Choose *Standard Property Search Manager.* |
| `Cache Stale Timeout` | No | If defined, determines the time in milliseconds after which old resources are deleted from the database.<br>The lifetime of cached resources is determined by the `Cache Timeout` parameter.<br>After the timeout, they are *stale*. A resource that is older than the specification in the `Cache Stale Timeout` parameter is removed from the cache. |

| | | |
|---|---|---|
| `Cache Timeout` | No | Timeout in milliseconds for resources in the cache. The timeout determines the amount of time for which cached resources (content and properties) are not refreshed. The optional value is determined by the update frequency of the remote site. However, the timeout value should be shorter than the update interval of the remote location, so that the cache is updated with the latest content. For resources supplied with an “Expires” header by the remote HTTP server, the timeout is added to the expiry date. |
| `HTTP Timeout` | No | Timeout in milliseconds after which operations on the server are aborted. The default value is 180000 ms (that is, three minutes). |
| **`Repository Services`** | No | Identifiers of the repository services you want to use with the repository. The `properties` repository service needs to be activated so that documents can be classified. |
| **`Web Sites`** | No | Selection of repositories stored in the root directory of the Web repository manager. |
| `ACL Manager Cache` | No | Cache identification for resource ACLs. This parameter is required if an ACL security manager is specified in the `Security Manager` parameter. You can choose the preconfigured cache *ca_rsrc_acl*. |
| **`Connection Pool`** | No | Identifier of a JDBC connection pool to be used by the Web repository manager for storing resource properties persistently in the database. This enables delta-crawling of the Web repository. You can use a memory cache and connection pool together. Choose *dbconrep* if you do not want to create your own JDBC connection pool. |
| **`HTML Property Extractors`** | No | Identifier of an HTML property extractor to be used with this repository manager for extracting additional resource properties. |
| **`Memory Cache`** | Yes | Identifier of the cache to be used by the Web repository manager for caching both content and properties of generated resources. |
| `Security Manager` | No | Specification of the security manager that controls access to repository contents. If you want the Web repository to perform an authorization check when resources are accessed, you need to specify a security manager. Generally, the *AclSecurityManager* should be used for Web repositories. |

Example Configuration for a Web Repository Manager




For more information on Web repositories, see the KM administration guide.

## 2.2 Configuring a Crawler Profile

This section of the How To guide describes how you create a new crawler profile.

Several crawler profiles are available in the system. However, by default, the number of document levels that are searched is limited (`Depth = -1`). A restriction is recommended in most cases since you otherwise have no control over the size of the Web repository.

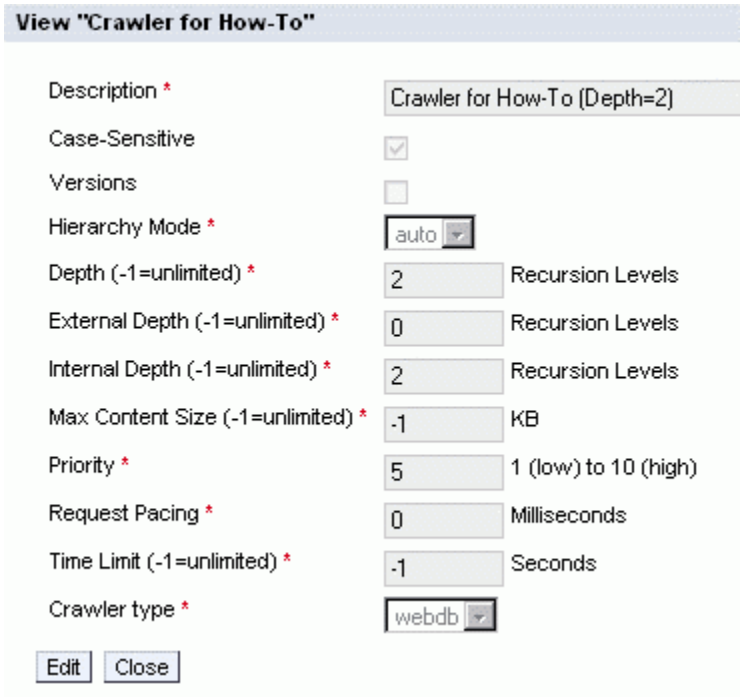
To create a new crawler profile, choose *Content Management* → *Global Services* → *Crawler Profiles.*

**Crawler Profile Parameters**

| Parameter | Required | Description |
|---|---|---|
| **`Name`** | Yes | Name of the crawler profile. |
| **`Description`** | Yes | Description of the crawler profile.<br>This description is used for the crawler profile listing in index administration. |
| `Case-Sensitive` | No | Specifies whether the crawler distinguishes between lowercase and uppercase for resources.<br>If it is activated, `/web/page/index.html` and `/web/page/Index.HTML` are handled differently, for example. |
| `Versions` | No | Boolean parameter that determines whether versions of a resource are included in the results set |
| **`Hierarchy Mode`** | Yes | Specifies how hierarchies are crawled.<br>Choose the setting `auto` for Web repositories. |
| **`Depth`** | Yes | Number of recursion levels applied in the crawling process.<br>For example, a recursion level of '`2`' means that starting from a given document, all documents referenced by hyperlinks in the start document and all documents referenced in turn in those documents are included in the results set.<br>In the case of a crawler that is used for a Web repository, the specification refers to links. The specification should not be less than the values in the parameters `External Depth` and `Internal Depth`.<br>Never choose –1 for a dynamic Web repository. |
| **`External Depth`** | Yes | Number of recursion levels used if the crawler finds links to documents in another repository/on another Web server.<br>If you enter 0, external links are not pursued.<br>The specification 1 is useful for Web repositories, since in this case links to further information on another server are pursued. |
| **`Internal Depth`** | Yes | Number of recursion levels used if the crawler finds links to documents in the same repository/on the same Web server.<br>If you enter 0, internal links are not pursued. |

| | | |
|---|---|---|
| `Max Content Size` | Yes | Specifies the maximum size (in KB) of the content of the resources that register with the TREX components.<br><br>**-1** means no limit. **0** means that only resources are registered, and not their content. This is useful if large documents (for example, pdfs), are not to be indexed. |
| `Priority` | Yes | Integer between 1 and `6` that denotes the priority of the crawler thread(s) on the portal server.<br><br>`Priority = 1` (lowest priority)<br>`Priority = 6` (highest priority) |
| `Request Pacing` | Yes | Integer that determines the size of time lags between successive requests the crawler sends to the Web server whose content is being crawled.<br><br>The default value is '0', meaning the crawler sends successive requests without any delay. |
| `Time Limit` | Yes | The time interval in seconds after which the crawler terminates the crawling process.<br><br>The default value is -1, meaning that no time limit is set. |
| **`Crawler Type`** | Yes | Specifies the crawler type.<br><br>Choose *web* or *webdb* to crawl a Web repository.<br><br>*web*: This crawler type is fast but requires a lot of memory.<br><br>*webdb*: This is slower than *web* but requires less memory since the crawled resources are stored in the database. Only this crawler type enables delta crawling. |

Example Configuration for a Crawler Profile




For more information on crawlers, see the KM administration guide.

## 2.3 Creating an Index and Assigning a Web RM

Carry out the following steps to crawl the Web repository for indexing. After the indexing process has been completed, you can start a search on documents that are stored in the Web repository.

The following steps are necessary:

1. Create an index.
2. Assign the Web repository to the index.
3. Select the crawler profile that is to be used to search the repository.
4. Define a plan for indexing the repository.

In EP 5.0, index administration is located in the *KM Admin* workset.

If you are using EP 6.0, it can be found under *System Administration* → *System Configuration* → *Knowledge Management*.

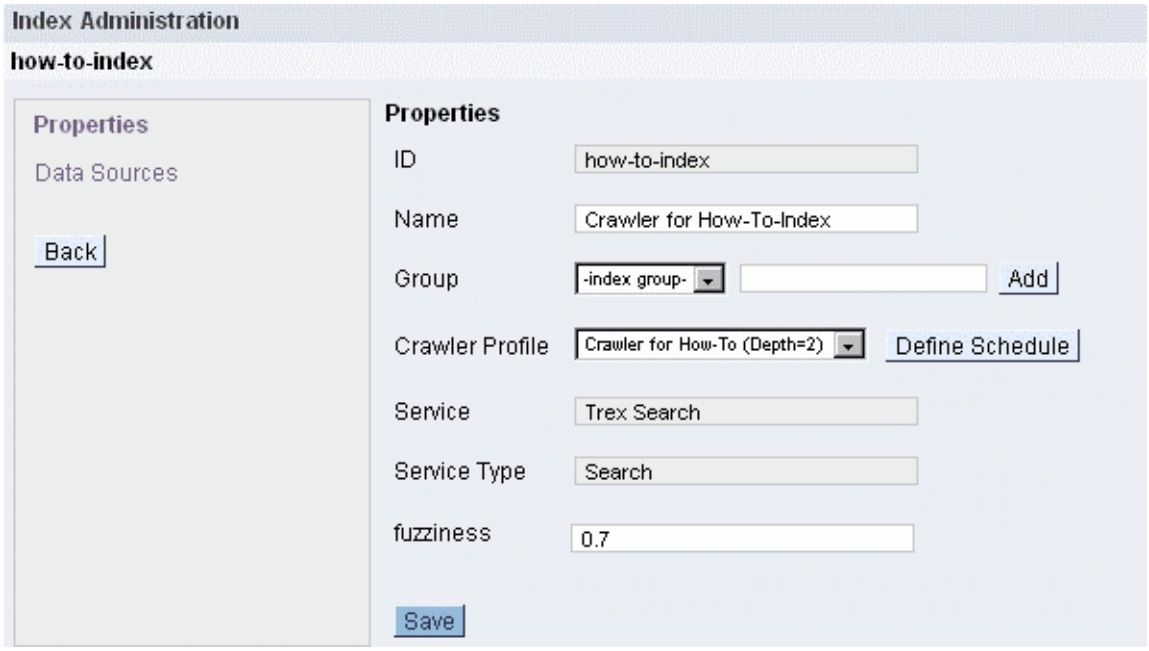
### 1. Create an Index

Choose *Create* in index administration.

Enter the required information and then save your entries.

| Field | Description |
|---|---|
| ***ID*** | Enter a unique ID for the index. |
| ***Name*** | Enter a name to be displayed in the search options later on (*Show Indexes*). |
| *Group* | An index group includes several indexes.<br>The name specified is displayed in the search options. |
| ***Crawler Profile*** | Select the crawler profile that you just created. |
| ***Service*** | Select an index service. In our example, a pure search index is used. If you want to use the Web repository in a taxonomy, choose a combined search and classification index. |

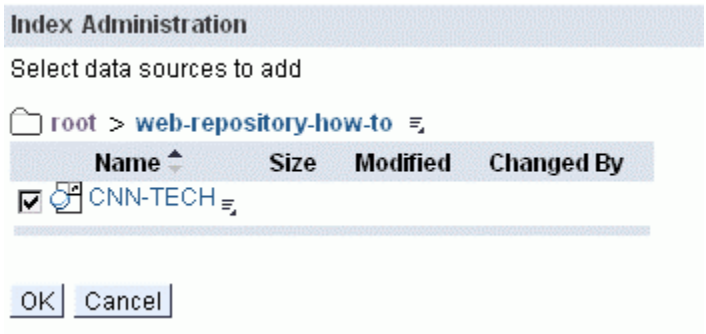
Example Configuration for an Index



### 2. Assign a Web Repository/Site to an Index

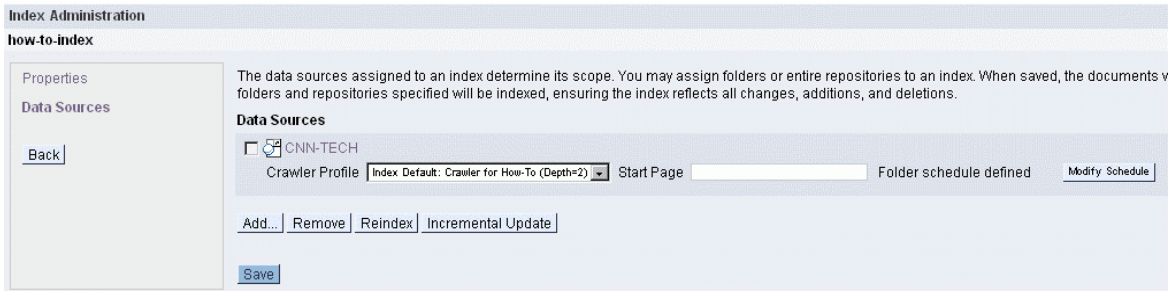
In the left-hand side of the index administration window, choose *Data Sources*. Navigate to the Web repository and select the Web site.

If the Web repository contains multiple Web sites, you can select the Web repository itself instead of the individual Web sites.



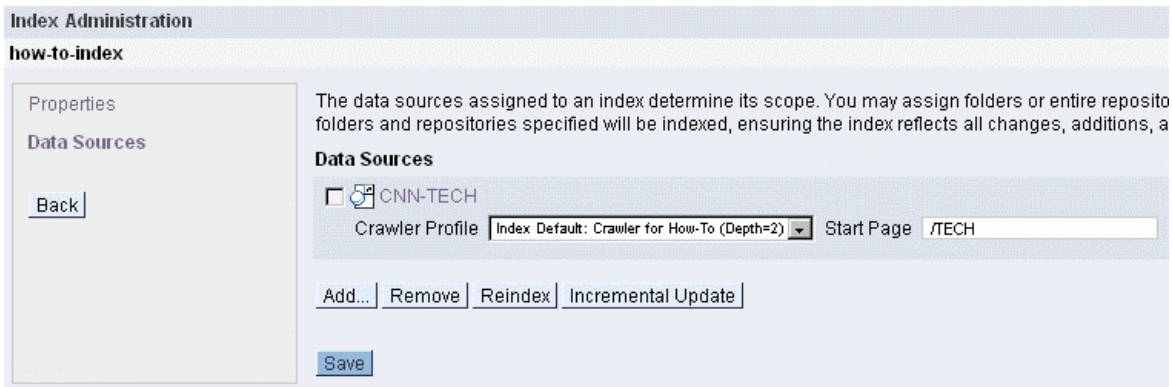

### 3. Select a Crawler Profile

Select the crawler profile that is to be used to search the Web site.




### 4. Specify the Start Page

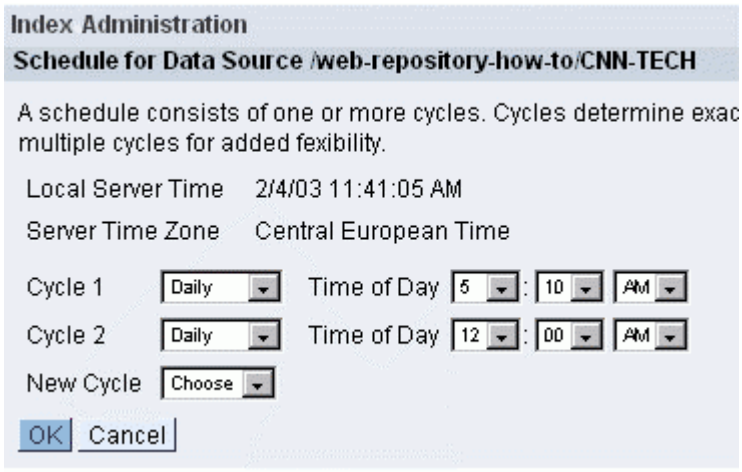
If you entered a start page in the definition of the Web site, you should also enter this start page here.



**5. Define a Plan for Searching and Indexing the Repository**

You can determine the times at which data sources are to be crawled. In our scenario, this takes place twice a day.

Call up the *Define Schedule* function and enter the required time schedule information.




For more information on indexes, see the KM administration guide.

## 2.4 Monitoring Crawling and Indexing

The KM platform allows you to monitor the crawling and indexing of Web repositories. You use the crawler monitor and TREX monitor to do this.

The crawler starts shortly after an index is created. You can monitor the status of indexing using the TREX monitor as soon as the crawling process has finished.

Note that the crawling and indexing processes can last several hours. The time required is dependent on the extent of the Web site.

In EP 5.0 the crawler monitor and TREX monitor can be found in the *KM Admin* workset.

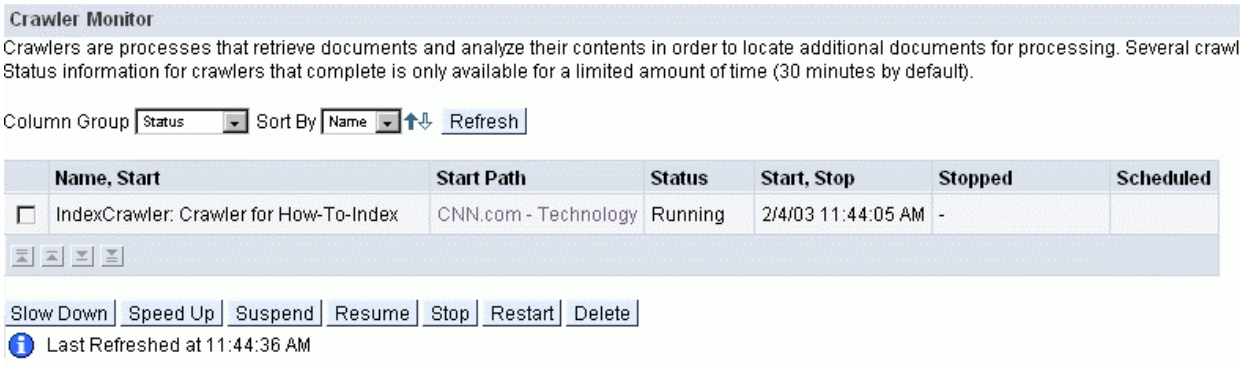
If you are using EP 6.0, they can be found under *System Administration* → *Monitoring* → *Knowledge Management*.

### Using the Crawler Monitor

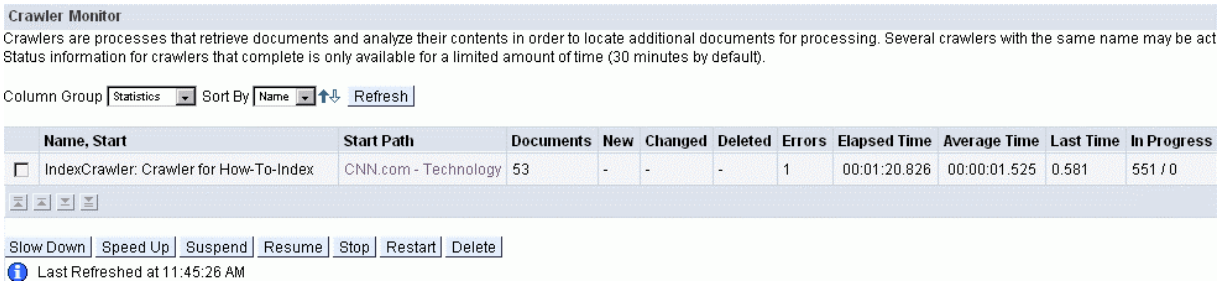
Call up the crawler monitor. You see an overview of the crawlers. This includes the crawler for crawling the Web repository CNN/TECH. The table below contains information on the status of the crawlers.

Status *Running*: The crawler has been started. The Web server is currently being searched.

Status *Completed*: The crawling process has been completed. Documents found are now indexed by TREX.




You can call up more statistics on the crawler. To do this, choose *Statistics* under *Column Group.* An overview shows you the number of documents found.




For more information on the crawler monitor, see the KM administration guide.

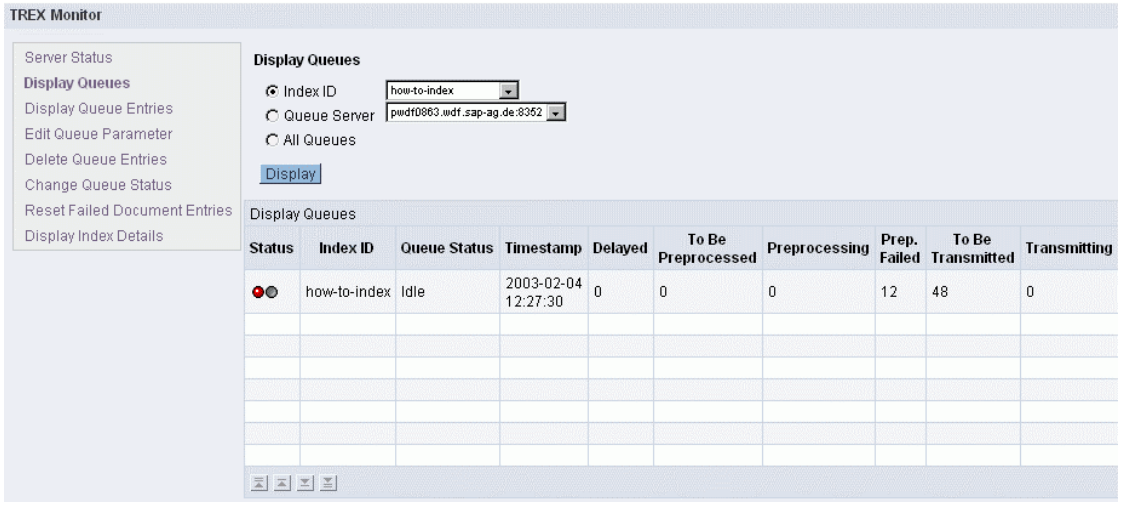
### Using the TREX Monitor

As soon as the crawling process has finished, you can check the current status of indexing using the TREX monitor.

### Displaying Queues for an Index

In the left-hand side of the window, choose *Display Queues*. Now select the index required in the *Index ID* field. You see an overview of the current processing status of all documents that are in the index queue.

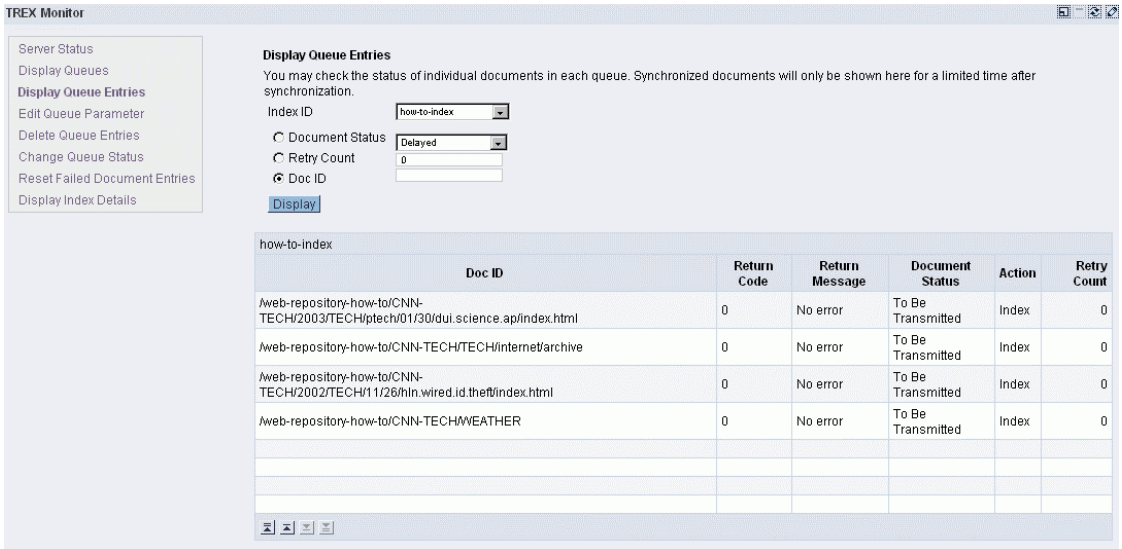
You see, for example, the number of documents that were already prepared for indexing.




### Displaying Entries in Queues

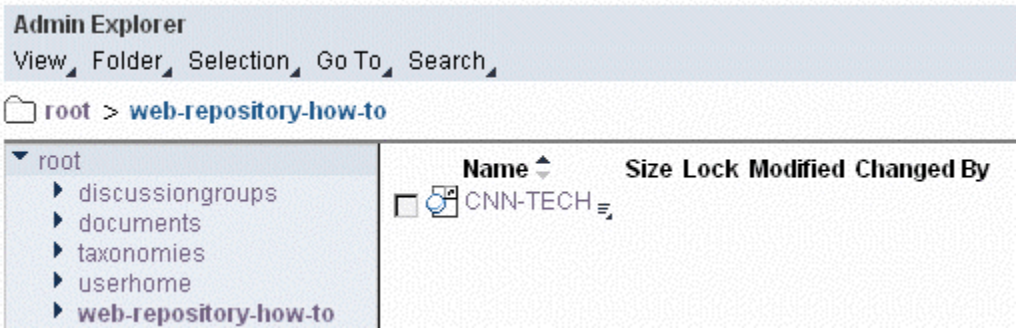
To see an overview of the files contained in an index queue, choose *Display Queue Entries* from the left-hand side of the screen.

The status *OK* means that the indexing of a document has been completed.



# 3 Result

Open the explorer of the Knowledge Management platform. You are now in the *web-repository-how-to* repository that contains the Web site *CNN-TECH*.




When you click on the Web site, it opens in a new window. You can now access the linked documents on the Web site.

Links that are invoked using Javascript cannot be opened.

When you carry out a search, documents from the CNN Web site are also listed in the results list.

# Copyright