

EXT: Real URL

Extension Key: **realurl**

Copyright 2003-2004, Martin Poelstra, <martin@beryllium.net>

This document is published under the Open Content License
available from <http://www.opencontent.org/opl.shtml>

The content of this document is related to TYPO3
- a GNU/GPL CMS/Framework available from www.typo3.com

Table of Contents

EXT: Real URL	1	FAQ.....	3
Introduction	1	Configuration	3
What does it do?.....	1	Installation.....	3
Features.....	2	Multi-language sites.....	5
Users manual	2	Reference.....	5
How does it work?.....	2	Known problems	5
Administration	3	To-Do list	6
		Changelog	6

Introduction

What does it do?

Typo3 works with page-IDs. This works great, however the URLs are very ugly. There are workarounds (simulateStaticDocuments), but that's just a fake: the ID must still be supplied in the URL, which is not desirable. Furthermore, only the page-title is shown, not the complete 'path' (or 'rootline') to the page.

Normally, you type in the path and filename of a document, but Typo3 works exclusively with page-IDs. The RealURL-extension provides a way to translate between page-IDs and (virtual) URLs that are easy to read and remember.

Typed URL	Typo3 id and type
http://www.domain.com/	id=0, type=0
http://www.domain.com/products/product1/features/	id=123, type=0
http://www.domain.com/products/product1/features/leftframe.html	id=123, type=2
http://www.domain.com/products/product1/features/	id=123, type=0

[Table 1 Single-language-site examples]

Typed URL	Typo3 id, type and language
http://www.domain.com/en/products/refrigerator/	id=456, type=0, langCode=en, langID=0
http://www.domain.com/nl/producten/koelkast/	id=456, type=0, langCode=nl, langID=2

[Table 2 Multi-language-site examples]

Although the extension works for me, it can seriously screw up your website if you configure it the wrong way, or if I made a little mistake somewhere, so use it at your own risk!

Features

- URLs are cached, so translating between URLs and IDs is very fast
- Pagetitles can contain spaces and characters like /.,&@ etc, the URL will still be nice.
- URLs are generated as nice-looking lowercase paths
- If a page is renamed, the old URL can still be used (see below in the Users Manual), so if the page was indexed by e.g. Google, it can still be found.
- It can handle different frames, or other pagetypes
- URLs are multilingual: if you're browsing in Dutch, you'll see Dutch URLs
- Once configured the systems works fully automatic, creating new and updating existing URLs
- You can easily see where shortcuts are pointing to, as the 'target' URL is generated, instead of the URL to the shortcut itself.
- It automatically handles multiple domains in the database
- It automatically handles installations of Typo in directories other than the root of the website too

Users manual

The extension works fully automatic once installed, so there should be no difference for Backend Users that create content elements etc. This section might be a good place to explain (in a pretty readable way) what it's doing, and how. Here we go:

How does it work?

The URLs generated by RealURL consist of 4 parts: a domainname, optionally a language, a path and a filename. The domainname and path determine the page-ID to be used. The language is looked up in an array and will tell which language to generate the page in. The filename is used to figure out which page-type is requested. This can be used to implement sites with frames.

Let's first take a look at the simple case, where everything goes smooth and nothing goes wrong:

When you type an URL (or click on it), it is looked up in the so called URL-cache. Assuming it is found, we then know the page-ID (from the path), pagetype (from the filename) and language (also from the path) to generate the correct page. And we're done.

Now, some things can go wrong here: first of all, the paths generated by RealURL only contain a..z, 0..9 and underscores ('_'), so it's a good idea to strip the URL off all unwanted characters before we look it up. So we do that :)

Furthermore, if the URL isn't found in the URL-cache (e.g. when the cache is cleared), we have to search for it in the database. This is done by first examining the domain-part, and then searching for every 'directory' in the URL until we reach the destination-page. The language-part (if present) is also translated to a language-ID. This result is cached, so we can use it lateron.

If the language wasn't given in the URL, a function is called to figure out what language will be most appropriate. I created some code which looks the IP-address up in the IP2Country-database (a table), which I imported into the Typo3-database. I might create a separate extension for this, but for now you can uncomment the code if you want to use it. Look for `getDefaultLangID()`.

When the page-ID is found, another check is done (after Typo3 has loaded all kinds of information about the page): it checks to see if the requested URL corresponds to the 'real' URL of the page. This might be different due to changes in the page-title of the page, or one of it's parents. Or, one might have typed the URL to a page that is a shortcut to the real page. In those cases, the user is redirected to the real/official/new URL of the page and in case of a changed page-title, the old URL is marked as 'expiring'.

This makes it possible to change the page-title of a page (and thus it's URL), but still be able to reach that page through the old URL, which will still be used by e.g. Google.

A problem arises when you create another page, with the same title as a page that existed before, because that URL still points to the other page. Therefore, if RealURL notices that you typed an expiring URL, it searches the database like the URL wasn't found in the cache. If a page is found this way, that new URL + page-ID is cached instead. If no page is found, the cached result will be used.

The other way around is much simpler: when an URL has to be generated for a certain page, it is looked up in the same URL-cache. If it isn't already there, the URL will be created by building the so-called RootLine for the page, filtering every page-title so that it contains only a..z, 0..9 and underscores and finally caching it. This process does take languages into account, so if you're browsing to the Dutch version of a page, you'll get a Dutch URL to it (using the Dutch page-titles).

Administration

As said above, the extension is self-maintaining, so there should be no (super-) user intervention necessary. However, because the underlying technology is pretty complex, it can happen that wrong information is cached for one or more URLs.

FAQ

First of all: see the Known Problems.

- Q: The wrong page is shown!

A: This is most likely caused by (a combination of) renaming a page (or one of its parents), deleting a page, creating a page with the same name as a page that existed before, etc.

The best way to try to solve the problem is to try request all pages from the root of the website all the way down to the problematic page itself. This should automatically remove old (changed) URLs. If this doesn't work, you must manually clear the URL-cache: go to the phpMyAdmin-module in the backend, select the tx_realurl_urllcache table and select the 'Empty'-tab to clear the table.
- Q: The browser is redirected to the same URL over and over again, in an endless loop. The page is never displayed!

A: See FAQ 1.
- Q: I just get a strange error on a blank page, like "url cache error (select from cache)". I don't get the real page.

A: This is probably caused by a misconfiguration.

Check if you can display a page as follows: <http://www.server.com/index.php?id=0&type=0>. This should give you at least the page. Now, the links in the page should be RealURLs. Try to see if the contents of the tx_realurl_urllcache table make sense to you (look it up in the phpMyAdmin-module).

Alternatively, you can try to set the debug-directive to 1 (see the Reference below) and try something like <http://www.server.com/some/page/titles/index.html?debug=10>. This will give you output in the browser-window, and in <http://www.server.com/typo3conf/debuglog.html>. Make sure the debuglog.html-file is writeable by the webserver for this to work.

Most likely you will see what is going wrong (wrong URLs are cached or something) by looking at the contents of the tx_realurl_urllcache-table.

If you want to go and trace execution in the code, look for these startingpoints:

```
function checkAlternativeldMethods: Finds the ID for an URL
function makeSimulFileName: Finds the URL for an ID
function fetch_the_id: Checks to see if the requested URL is the 'official' URL for the page, this one generates the redirects.
```
- Q: Something is wrong with the conversion between pagetypes and filenames.

A: Make sure you DON'T include the extension with typeToPage, but DO include the extension with fileNameToType. See the reference and configuration examples.
- Q: Some images / javascript aren't showing up / aren't working.

A: You forgot to modify your templates to provide an absolute base. See Step 3 in the Configuration.
- Q: How do set up an installation of Typo3 other then in the document-root?

A: Create a domain-record like www.server.com/path_to_typo in the root of that website. Please make sure the "Is Siteroot"-box is checked on the root-page.
- Q: RealURL doesn't recognize my domains!

A: Create the appropriate domain-records on the pages and check the "Is Siteroot"-box of those pages.

Configuration

Installation

To install this extension, four steps must be taken:

1. Install it in the Extension Manager
2. Configure Apache
3. Modify your templates for Real URLs
4. Configure the extension in typo3conf/localconf.php

Install the extension

This is documented very well in the usual Typo docs: just click the little gray sphere with the plus-sign and when it asks for any changes to commit, let it make them. It's not doing anything yet though.

Configure Apache

RealURLs work by providing 'virtual paths' to 'virtual files'. These don't actually exist on the file-system, so you must tell Apache to let a PHP-script handle the request if it can't find the file. This way, all URLs to pages (like www.server.com/products/product1/left.html) will be 'redirected' to /index.php, which will try to find the path in the URL-

database of RealURLs. Real files (like images, the Typo3 backend, static html-files, etc.) will still be handled by Apache itself though.

You should put the supplied sample .htaccess file (called `_.htaccess`) in the root of your Typo3-installation.

Alternatively, you could include the following lines in your `httpd.conf`, probably in the VirtualHost-section:

```
RewriteEngine On
RewriteRule ^/typo3$ - [L]
RewriteRule ^/typo3/.*$ - [L]

RewriteCond %{DOCUMENT_ROOT}%{REQUEST_FILENAME} !-f
RewriteCond %{DOCUMENT_ROOT}%{REQUEST_FILENAME} !-d
RewriteCond %{DOCUMENT_ROOT}%{REQUEST_FILENAME} !-l
RewriteRule .* /index.php
```

This will tell Apache that it should rewrite every URL that's not a filename, directory or symlink. It leaves everything starting with `/typo3/` alone too.

Modify your templates

By default, Typo3 generates all links to other pages as www.server.com/index.php?id=123&type=0, so all pages seem to be in one (filesystem-) directory: the root of the website. The problem is, that many extensions (and Typo3 core code) rely on images, javascripts, etc. to be in a directory relative to the Typo3-root, like `typo3/ext/indexed_search/pi/res/pages.gif`. This approach doesn't work when the path is constantly changing. There is a TypoScript-setup directive to set an absolute prefix to all links and images (`config.absRefPrefix`), but sadly enough that isn't implemented in all places (the indexed-search and front-end-editing for example), so that doesn't work too well.

There is a very simple solution in HTML though: just supply the `<base>`-tag in the `<header>` of your pages, like:

```
<BASE href="http://your.domain.com/">
```

To make your TypoScript templates RealURL-enabled, you should therefore include this statement in your HTML-templates, or use the following TypoScript snippet:

```
page.headerData.1 = HTML
page.headerData.1.value = <BASE href="http://your.domain.com/">
```

Please don't use `config.absRefPrefix`. It has some nasty properties that render RealURLs complete unusable sometimes. The only problem is that the 404-page of Typo doesn't have the `<base>`-tag, so it doesn't show the Typo3-logo :)

Make sure you include it in ALL page-types that are generated!

Now, you should create Domain-records on the pages where domains start. Even if you only have one domain, it's a good idea to create a Domain-record for it. There's one thing you should note:

If you have installed Typo3 in the document-root of a host, you should create a domain-record named like www.server.com. If, on the other hand, your Typo3-installation is in a different directory, you should create a domain-record named something like www.server.com/the_path_to_your_typo. Slashes at the end don't matter that much.

Configure the extension

Now, the only thing left to do is configure the extension itself. I've included two files called `localconf.simple.php` and `localconf.advanced.php`, from which you can copy an example configuration. You should copy ONE of these configuration-samples to (the beginning of) your `typo3conf/localconf.php`. If you have a site with frames and/or languages, you should also adjust the variables to your needs.

Clear the cache once more (including the cached files in `typo3conf/`) to be sure there's no old stuff left.

For your convenience, here are two examples with a little explanation:

Example configuration: the simple version (localconf.simple.php)

```
////////////////////////////////////
// Simple setup:
// - no frames or other pagetypes
// - no languages
// - this works pretty well with the Typo3 Testsite-package
$TYPO3_CONF_VARS["FE"]["realurl"]["enabled"] = '1';
$TYPO3_CONF_VARS["FE"]["realurl"]["expireDays"] = 60; // Value is in days
$TYPO3_CONF_VARS["FE"]["realurl"]["debug"] = 0; // Set this to 0 (zero) on production-servers!
$TYPO3_CONF_VARS["FE"]["realurl"]["langCodes"] = array(0 => ''); // Leave empty to disable a language-
prefix in the URL
$TYPO3_CONF_VARS["FE"]["realurl"]["langDescs"] = array(0 => '');
$TYPO3_CONF_VARS["FE"]["realurl"]["typeToPage"] = array(0 => 'index', 98 => 'printer');
$TYPO3_CONF_VARS["FE"]["realurl"]["fileNameToType"] = array('' => 0, 'index.html' => 0, 'printer.html'
=> 98);
```

Example configuration: the advanced version (localconf.advanced.php)

```
////////////////////////////////////
// Advanced setup:
// - 3 frames (frameset = typeNum 0, page = 1, menu = 2, top = 3) + printerfriendly-version (=98)
```

```
// - 2 languages: English (id 0) & Dutch (id 2)
$TYPO3_CONF_VARS["FE"]["realurl"]["enabled"] = '1';
$TYPO3_CONF_VARS["FE"]["realurl"]["expireDays"] = 60; // Value is in days
$TYPO3_CONF_VARS["FE"]["realurl"]["debug"] = 0; // Set this to 0 (zero) on production-servers!
$TYPO3_CONF_VARS["FE"]["realurl"]["langCodes"] = array(0 => 'en', 2 => 'nl');
$TYPO3_CONF_VARS["FE"]["realurl"]["langDescs"] = array(0 => 'English', 2 => 'Nederlands');
$TYPO3_CONF_VARS["FE"]["realurl"]["typeToPage"] = array(0 => 'index', 1 => 'page', 2 => 'menu', 3 =>
'top', 98 => 'print');
$TYPO3_CONF_VARS["FE"]["realurl"]["fileNameToType"] = array('' => 0, 'index.html' => 0, 'page.html' =>
1, 'menu.html' => 2, 'top.html' => 3, 'print.html' => 98);
```

Multi-language sites

Languages are handled by prepending the path with the 2-character languagecode, like:
<http://your.domain.com/en/services/consultancy/index.html>

Currently there's no way to change this to the 'file-extension-method' (like `index.html.en`). To use the requested language in your pages, RealURL simulates the default 'getvar'-behaviour: it will create a variable named L, just like you used to have in e.g. `http://your.server.com/index.php?id=123&type=0&L=2`

There is also another way: it will also set the language-ID and the 2-character abbreviation of the country-code on the \$TSFE-object (called \$TSFE->langID and \$TSFE->langCode respectively). You can use these properties like this (put it in the Setup of your TypoScript Template):

```
# English (or otherwise the default language)
config.sys_language_uid = 0
config.language = en
temp.aString = The English version of the string

# Dutch (nl)
[globalString= TSFE:langCode=nl]
config.sys_language_uid = 2
config.language = nl
temp.aString = De Nederlandse versie van de bovenstaande string
```

Reference

Below you'll find an overview of all configuration directives and their meaning.

Directive	Description
enabled	If set to 1, it will translate to and from RealURLs. Set to 0 to disable. Don't forget to clear the cache...
expireDays	The time the old URL of a page whose pagetitle changed will still be remembered (in days)
debug	If you set this to 1, and supply e.g. <code>?debug=10</code> in the URL, you'll get debug-output in the page and in <code>/typo3conf/debuglog.html</code>
langCodes	Used to translate between the 2-character language-codes and the Typo3 language-id as defined by Language-records. If you don't use languages, set it to an empty array.
langDescs	Not really used in RealURLs, but can be used in e.g. a modified version of the indexed search-engine to display the name of a language. The problem is that the default language isn't defined in the <code>sys_language</code> -table.
typeToPage	Converts a page-type to a part of the filename. The <code>.html</code> -extension is added by some code in Typo and can't (easily) be changed :(So don't set type 0 to an empty string, it will yield <code>.html</code> as the filename! Don't include the <code>".html"</code> !
fileNameToType	Translates filenames back to a type. You can create several filenames that translate to the same pagetype if you want to. DO include <code>".html"</code> !
spaceCharacter	Normally, this defaults to an underscore (<code>_</code>), which is used to replace spaces and such in an URL. You can set this to e.g. a hyphen (<code>-</code>) if you want to.

[Table 3 RealURL Configuration directives]

Known problems

- Languages are currently only handled by prepending the path with the language-code. I might include a feature to be able to produce urls like `index.html.nl` too. Let me know if you need it.
- The link to the printerfriendly version page in the TestSite-package doesn't work correctly.
- It might give problems with extensions passing variables through the URL, haven't tested this yet.
- Links to pages with type 'URL' will be generated as though they were normal pages. The link where they point to should be taken actually.
- MD5 and Base64 parameters don't work yet (I think).
- Pages that have subpages with the same page-title (actually: that resolve to the same URL) don't work yet. See TODO.

To-Do list

- Include information on how to set this up with IIS, if possible at all. Can somebody please try this?
- Extensive testing still has to be done, especially with changing page-titles, deleting a page, then creating another page with the same name, the multiple-domain-stuff, etc.
- Some clever 404 page can be created, using the builtin indexed-search for example. The requested language (/nl/...) could help to provide the page in the requested language. Nothing done with this yet though.
- There's a function to find the desired language when it's not explicitly given in the URL. It needs some more work to be generally useable though. (Take a look at the source, there's code to figure out the language with the IP2Country database...)
- The table should be automatically cleaned up once in a while, to remove expired URLs. Haven't looked into this yet. One possibility is to create a cron-job for it, another way is to let a random number decide that it is time to delete all old URLs in the function updateURLCache()
- Implement a way to generate nice URLs to things like posts in a forum, or a news-article.
- I could create an option to enforce usage of the https-protocol quite easily.
- Remove all calls to the status()-function when the thing gets a little more stable.
- Make configuration somewhat easier, so no localconf.php-tweaking.
- Thinking of a way on how to let extensions profit from RealURLs (like newsarticles and forum-posts).
- Making the pathPrefix-detection (the path/to/typo/) more elegant.
- Fix bugs and rest of TODO-items in source (denoted by // !! TODO !! Bladiebla).

Changelog

- 0.0.1: First upload
- 0.0.4: First working version
- 0.0.5: An icon was added, thanks to Netcreaters for creating it!
- 0.0.6: Documentation in StarOffice format added
- 0.0.7: Almost complete rewrite / revision of the code, implemented '/path_to_typo'-feature, implemented support for multiple domains, changed the code so that most of the configuration is now automatic, updated documentation
- 0.1.0: First publicly available version
- 0.1.1: pathPrefix didn't work correctly, so a hack was added to allow it too work now
- 0.1.2: URLs in pages weren't rendered correctly on single-language-sites
- 0.1.3: Added possibility to choose another character to replace a space (instead of an underscore), fixed another stupid bug regarding the rendering of some URLs