

Searching & Indexing in Plone/Zope

by Wouter Vanden Hove

Submitted for the Course of
Document Management,

KIM 2003-2004

Table of Contents

1. Introduction.....	2
2. Search-forms in the User-Interface.....	2
3. Search Engine Back-End.....	4
3.1 ZCatalog.....	5
3.2 Indexing.....	5
3.3 Full-text Indexing.....	7
4. Switching from ZCTextIndex to TextIndexNG.....	12
5. Empirical Comparison of TextIndex and TextIndexNG2.....	15
6. Our Optimal Configuration	22
7. Thesaurus.....	24
8. Integrating Office Documents in Plone CMS.....	25
9. Other Open Source Indexing Software.....	26
10. Conclusion.....	27
11. References.....	28

1. Introduction

In the first paper “Creating Multilingual Websites with Plone/Zope”, we created a multilingual website in seven languages. We translated the interface of an additional Product, PloneSearchBox. But we didn't actually examine this search-utility more closer.

What happens if you submit a search-query? What are its possibilities? Does it just use Google as backend, like so many other websites? The answer to that last question is clearly “no”, because Zope has its own configurable and extensible search-engine.

In this paper we study in detail the search-engine underlying the Plone Content Management system. We touch upon various concepts that are important, like indexing, splitting, stemming,... and show how they affect search-results in practice. In practice, we compare the default indexing-system called “ZCTextIndex” with the more advanced “TextIndexNG”.

Since both search-engine backends cannot be used at the same time, and take non-negligible reconfiguration to switch, we initiated a second version of our multilingual site. They can be found at:

- http://minfpc26.vub.ac.be:9080/multilingual_site (ZCTextIndex)
- http://minfpc26.vub.ac.be:8080/multilingual_site (TextIndexNG)

Since we have created a truly multilingual website, the search-technology will need to encompass this. Indexing, splitting words, leaving out stop words, are very language dependent. Where possible, we point to solutions that can handle searching multilingual websites.

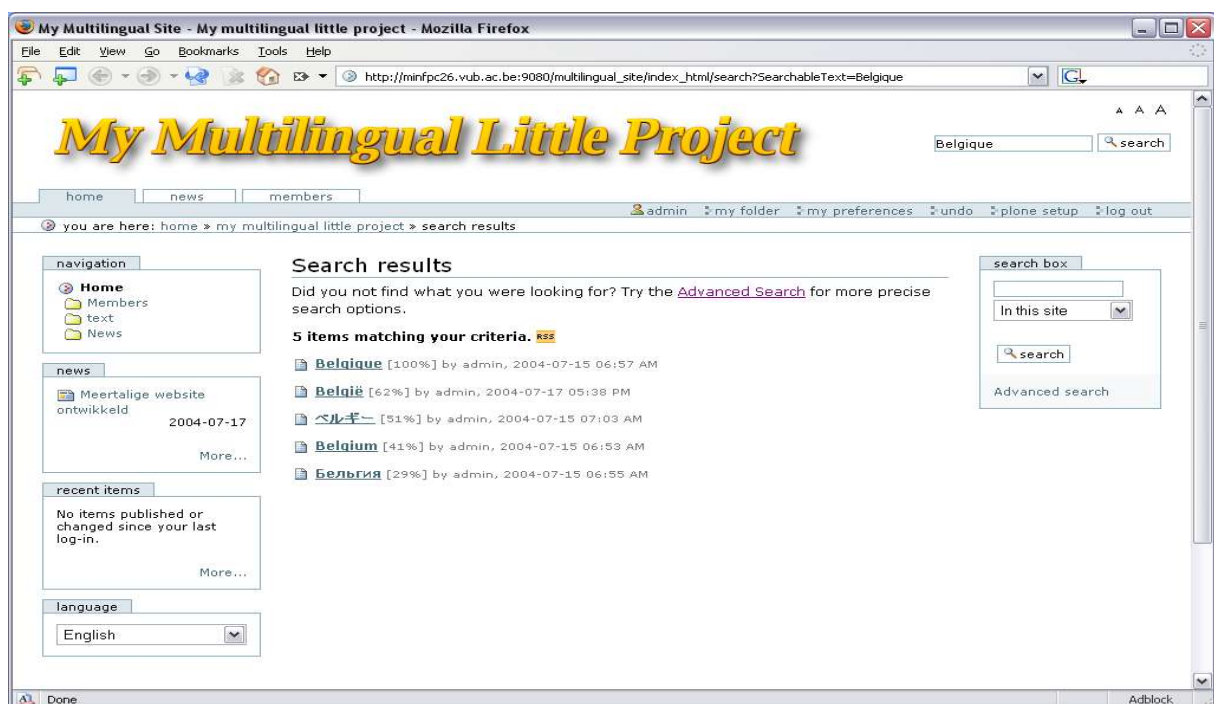
2. Search-forms in the User-Interface

By default Plone features two search-forms: a simple search-box and a more advanced search-form. These are the only things the end-user sees when using the search-engine.

2.1 Simple Search

In the top-right of the website, there is a little search-box by default. In the previous paper, we also installed an additional module called “PloneSearchBox”. Both are shown in figure 1.

Try to give in some terms in the searchbox, even multilingual terms. In figure 1, we see the results of a search-query for the terms “Belgique”.



Afbeelding 1 Two simple simple forms on the right

The little orange RSS-logo, seen in figure 1, provides a links to have these result in RSS-format: http://minfpc26.vub.ac.be:9080/multilingual_site/index_html/search_rss?SearchableText=Belgique. For more advanced users, this can be a very handy shortcut, to all recent papers, or dicuments of a certain author or department, or to documents with certain metadata or keywords.

The query-results are not just mere links, but have a parameter containing the search-term itself: “/multilingual_site/Text/belgium/nl/view?searchterm=Belgique” instead of merely “/multilingual_site/Text/belgium/nl/view”. The additional parameter is needed to highlight the search-term in the document: very similar to Google, it highlights the search-term in a bright highlighted color.



Afbeelding 2 SearchTerm “Belgique” highlighted in the document

2.2 Advanced Search-form

The search-box shows a link to a more advanced search-form, shown (partly) in figure 3.

The advanced search forms allows for searching very specific information, based on the Dublin core Metadata¹ and workflow status. There are search-boxes for searching based on: search text , title, keywords, description, new items since , item type , author review status,...

Experience shows that complex forms as this one, are not used very much by visitors. But in very large and complex websites with many thousands of documents, these advanced search-features are very necessary.

¹ <http://dublincore.org>

Description
Return items matching this description. Multiple words may be found by combining them with **AND** and **OR**.

New items since
Return items added since you were last logged on, the last week, etc.

Item type
Return items of a specific type.

Select All/None

Discussion Item

Document

Event

Favorite

File

Folder

I18NFolder

I18NLayer

Image

Large Plone Folder

Link

News Item

Plone Site

TempFolder

Topic

Author
Return items created by a particular person.

Review status
As a reviewer, you may search for items based on their review state. If you wish to constrain results to items in certain states, select them from this list.

Select All/None

published

visible

Afbeelding 3 Bottom part of the advanced search form

3. Search Engine Back-End

We covered the search-forms and the result ordinary visitors see. But the interesting part of a search-engine lies of course in the backend, shielded from normal user's eyes.

Normally one searches for a string of text inside a document. But Plone is build on top of a rather complex framework with workflows and metadata, so an efficient search-engine should make use of that.

If you look for a document written by a certain author, it is better to search in the meta-data fields, then in the full text of all documents. The advanced forms allows for that. But how does it work?

The answer is that Plone automatically indexes the entire website with many different indexes that all index separate parts of the available documents.

The important tool responsible for the searching and indexing in Plone is the **portal_catalog**. Figure 4 show the structure of this tool. This contains a vocabulary and/or lexicon,

3.1 ZCatalog

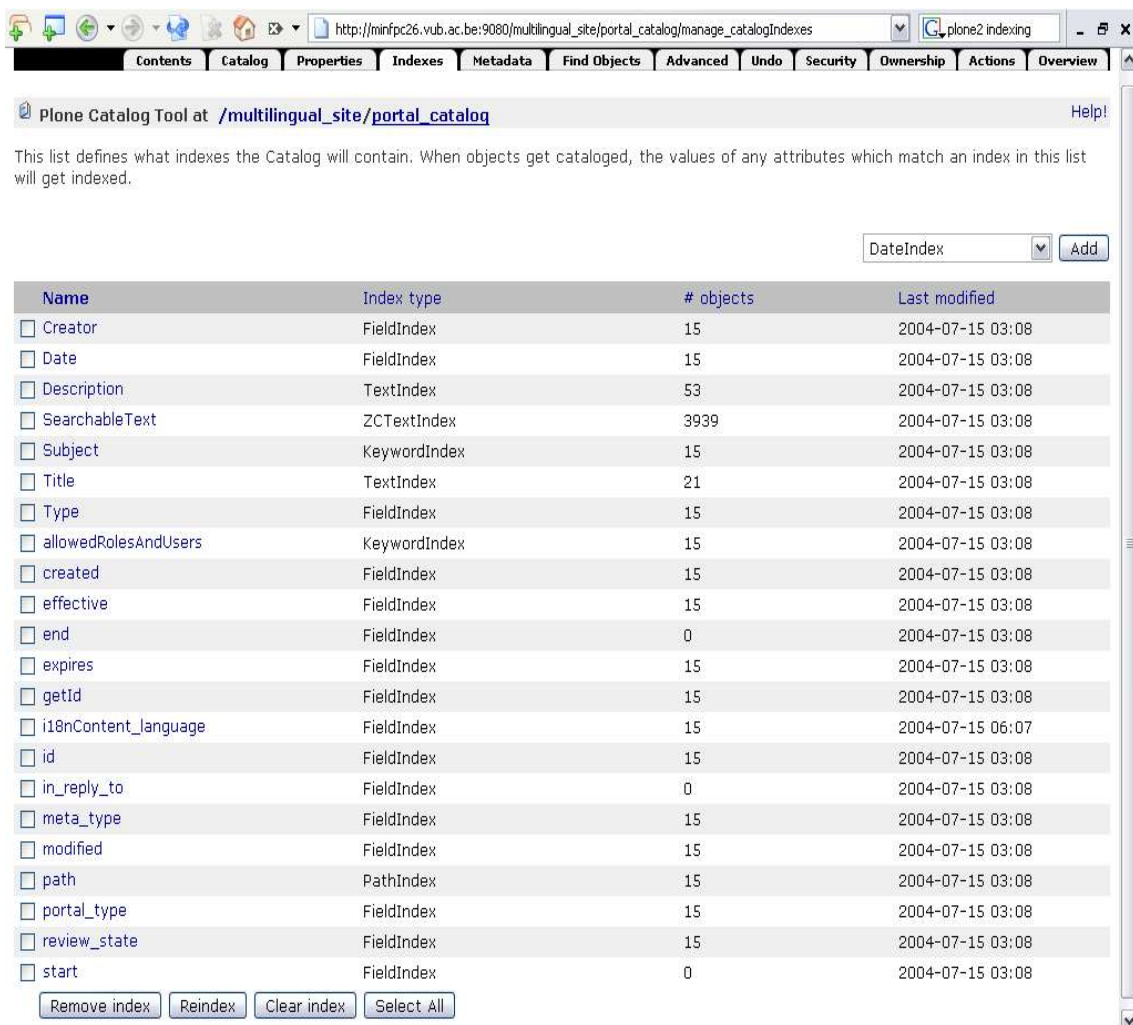
The ZCatalog (**Z**ope **C**atalog) provides powerful indexing and searching on a Zope database using a Zope management interface. A ZCatalog is a Zope object that can be added to a Folder, managed through the web, and extended in many ways.

ZCatalog is a Zope-aware wrapper around Catalog that by itself can be used outside the Zope framework. The only requirement for using Catalog is that you are using ZODB as your object store.

- Download & Install : <http://cvs.zope.org/Products/ZCatalog>
- Overview : <http://www.zopewiki.org/ZCatalog>
- Z Catalog Tutorial: <http://www.zope.org/Documentation/How-To/ZCatalogTutorial>
- Searching And Categorizing Content : <http://zopewiki.org/ZopeBookChapter17SearchingAndCategorizingContent>
- Advanced ZCatalog Searching: <http://www.zope.org/Members/Zen/howto/AdvZCatalogSearching>

3.2 Indexing

Plone comes with a wide variety of indexers. Several additional indexing-products can be downloaded & installed. We will review several of them. Figure 4 shows the default indexers used to index our multilingual plone-site.



The screenshot shows the 'manage_catalogIndexes' page in the Plone management interface. The browser address bar shows the URL: `http://minipc26.vub.ac.be:9080/multilingual_site/portal_catalog/manage_catalogIndexes`. The page title is 'Plone Catalog Tool at /multilingual_site/portal_catalog'. Below the title, there is a description: 'This list defines what indexes the Catalog will contain. When objects get cataloged, the values of any attributes which match an index in this list will get indexed.' There is a search box containing 'DateIndex' and an 'Add' button. The main content is a table with the following columns: Name, Index type, # objects, and Last modified. At the bottom of the table, there are four buttons: 'Remove index', 'Reindex', 'Clear index', and 'Select All'.

Name	Index type	# objects	Last modified
<input type="checkbox"/> Creator	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> Date	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> Description	TextIndex	53	2004-07-15 03:08
<input type="checkbox"/> SearchableText	ZCTextIndex	3939	2004-07-15 03:08
<input type="checkbox"/> Subject	KeywordIndex	15	2004-07-15 03:08
<input type="checkbox"/> Title	TextIndex	21	2004-07-15 03:08
<input type="checkbox"/> Type	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> allowedRolesAndUsers	KeywordIndex	15	2004-07-15 03:08
<input type="checkbox"/> created	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> effective	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> end	FieldIndex	0	2004-07-15 03:08
<input type="checkbox"/> expires	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> getId	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> i18nContent_language	FieldIndex	15	2004-07-15 06:07
<input type="checkbox"/> id	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> in_reply_to	FieldIndex	0	2004-07-15 03:08
<input type="checkbox"/> meta_type	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> modified	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> path	PathIndex	15	2004-07-15 03:08
<input type="checkbox"/> portal_type	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> review_state	FieldIndex	15	2004-07-15 03:08
<input type="checkbox"/> start	FieldIndex	0	2004-07-15 03:08

Afbeelding 4 22 different Indexers in the portal_catalog

Description of those indexers

- **Date Index** : indexes DateTime attributes.
- **Date Range Index** : A DateRangeIndex takes the name of two input attributes; “Since field” containing the start date of the range, and the “Until field” containing the end of the range. This index is filled with range information based on those two markers. You can then search for objects for those where a given date falls within the range.
- **Field Index** : Field Indexes treat the value of an objects attributes atomically, and can be used, for example, to track only a certain subset of object values, such as “meta_type”.
- **Keyword Index** : Keyword Indexes index a sequence of objects that act as 'keywords' for an object. A Keyword Index will return any objects that have one or more keywords specified in a search query.
- **Path Index** : A PathIndex indexes the physical path of all objects inside a catalog. It allows you to search for objects beginning or containing a special path component or a set of path component. A path component is defined as /<component1>/<component2>/.../<object_id> . Note: the “object_id” will not be indexed by the PathIndex.
- **Topic Index** : A TopicIndex is a container for so-called FilteredSets that consist of an expression and a set of internal ZCatalog document identifiers that fulfill this expression. TopicIndexes are useful for performance reasons when search queries take too long and pre-calculated result sets offer a better performance.

3.3 Full-text Indexing

Text Indexes break text up into individual words, and are often referred to as full-text indexes. Text indexes sort results by score meaning they return hits in order from the most relevant to the least relevant.

3.3.1. TextIndex

TextIndex is now deprecated but it is still present in Plone, just as the used vocabulary. It has been replaced by ZCTextIndex.

3.3.2. ZCTextIndex

ZCTextIndex (Zope Catalog TextIndex, <http://www.zopewiki.org/ZCTextIndex>) is the current built-in full-text index for Zope ZCatalog. It replaces TextIndex as of Zope 2.6.0. It is shipped with Plone2, therefore it's the default indexer in our multilingual site. It supports features like advanced relevance ranking, globbing support, boolean operators, phrase matching and a pluggable lexicon which is extensible to add additional text-processing features. We describe these features in more detail.

A ZCTextIndex Lexicon processes and stores the words of documents indexed with a ZCTextIndex. Multiple ZCTextIndexes can share the same lexicon.

1. Pipeline

Source text to be indexed, passes through a pipeline of processors that can effect the text indexed in various ways. At a minimum, the source text is passed through a splitter which divides the text on word boundaries. Query text is also processed by the same pipeline. The pipeline is configured for a lexicon when it is created.

- **Splitter:** Two splitters are available with Zope, a simple whitespace splitter, and an HTML-aware splitter which removes HTML markup from the source text while splitting. There are several specialised splitters for asian languages to replace the default splitters in ZCTextIndex or to complement them. This is because Japanese words for example are normally not delimited by spaces as in English. This page <http://www005.upp.so-net.ne.jp/nakagami/tips/ZCTextIndex.html> (unfortunately only in Japanese) mentions many different splitters: **ejSplitter**, **ChaSplitter**, **JSplitter**, **AJSplitter**, **MecabSplitter**, **JUTSplitter**, **ZCJPSplitter**.
 - JSplitter provides a replacement for the word Splitter that ZCatalog uses, allowing ZCatalog to do full text search of Japanese. JSplitter uses the freely available ChaSen library to do morphological analysis on Japanese text, which informs the breaking of the text into words.
 - CJKSplitter is a ZCTextIndex splitter for CJK (Chinese-Japanese-Korea) text stored as Unicode.
- **Case Normalizer:** Enable this for case-insensitive searches. Disable for case sensitive searches.
- **Stop Word Remover:** Removes common English words like articles, which are typically not helpful in narrowing searches and match most or all documents which reduces search speed. You can also choose to remove single letter words.

2. Relevance Ranking

There are two algorithms for scoring documents with ZCTextIndex:

- **Okapi BM25 Rank:** This is the best general purpose choice when the search text is much shorter than the text index. More information about BM25-Ranking can be found in the literature [8].

- **Cosine Rule:** This is the classic algorithm from the Manage Gigabytes book [3]. It works best when the query length is close to the length of the documents indexed. This makes it useful for finding documents that are "similar" to another document.

3. Boolean Operators

- **and:** document must match both terms on each side (the default if no operator is specified).
- **or:** document may match on either term on each side.
- **and not:** document matches the term on the left but not the term on the right. Note that **not** alone is not a legal operator.

Terms may be single words, phrases or multi-word operations in parenthesis.

4. Phrase Matching

Phrases are denoted by placing the words in double quotes. A phrase match also results if the words are separated by punctuation, but not space (e.g. 2.7 or 10,000).

3.3.3. FieldedTextIndex

This is an indexing product that is not present in Plone by default. It is available as a separate download from <http://zope.org/Members/Caseman/FieldedTextIndex>. FieldedTextIndex is a full-text indexing product based on ZCTextIndex for content with multiple textual fields. It allows indexing and searching any number of arbitrary text fields using a single index. As this explanation is straight-forward and there are no other differences with ZCTextIndex, we decide not to explore this product further.

3.3.4. TextIndexNG

TextIndexNG (NG=**N**ew **G**eneration) is the new full-text index for Zope and is the most feature-complete solution for fulltext indexing under Zope. The current stable release is 2.0.8. Between 2.0 and 2.0.8 the name was TextIndexNG2, but the 2 is removed has been dropped as of 2.0.8.

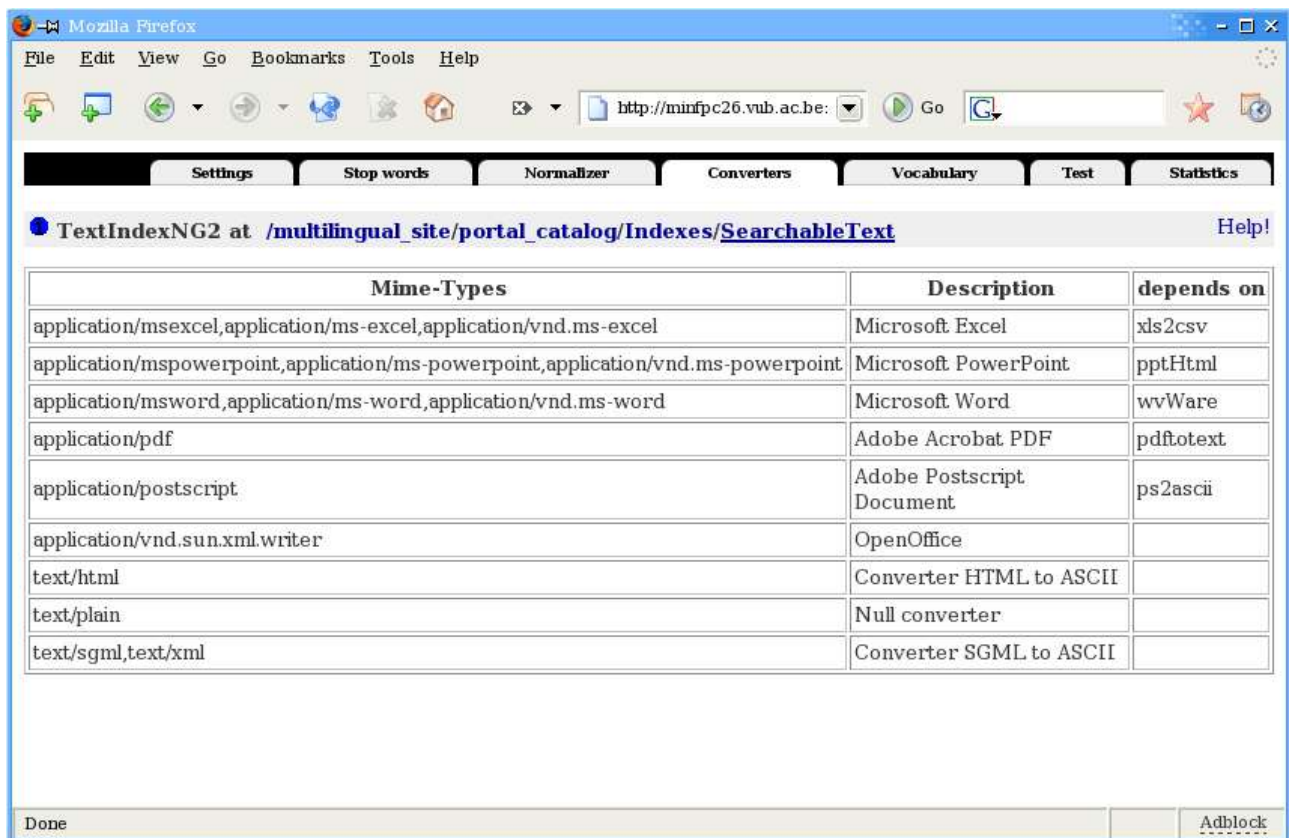
1. Download & Install : <http://www.zope.org/Members/ajung/TextIndexNG>
2. Documentation : <http://zope.org/Members/ajung/TextIndexNG/wiki/TextIndexNG>

3.3.4.1. DocumentConverters

TextIndexNG supports a registry for external converters wrapped into a Python class to convert a document or an object to text before it gets passed to the splitter. The converter is selected based on the mime-type and the extension of the object. Supported formats are

- HTML, PDF (requires xpdf:)<http://www.foolabs.com/xpdf>
- Postscript (requires ghostscript: <http://www.cs.wisc.edu/~ghost>)
- WinWord (requires wwWare:<http://wwware.sourceforge.net>),
- PowerPoint (requires pphtml)
- OpenOffice : www.openoffice.org
- All other converters from the DocumentLibrary product.

Figure 5 shows these installed converters.



Afbeelding 5 Available converters for indexing documents in other fileformats.

3.3.4.2. Stemming

A **stemmer** is a program or algorithm which determines the morphological root of a given word. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems. For example the morphological root of “talking” and “talked” is talk.

For efficient indexing of an entire website, only stemmed word or morphological roots are put into the index. A search-engine with stemming-support changes the inputted search-query into its morphological root, and then searches for this word in the index.

TextIndexNG has built-in stemmer support for the following languages: Danish, Dutch, English (Porter), Finnish, French, German, Italian, Norwegian, Portuguese, Russian, Spanish, Swedish. This stemmer support is built on top the Snowball² project that has an online stemmer-demonstration available at: <http://snowball.tartarus.org/demo.php>

The English stemming algorithm is better known as the **Porter Stemming Algorithm**. Detailed information about this algorithm, including several implementations can be found at <http://www.tartarus.org/~martin/PorterStemmer/index.html>

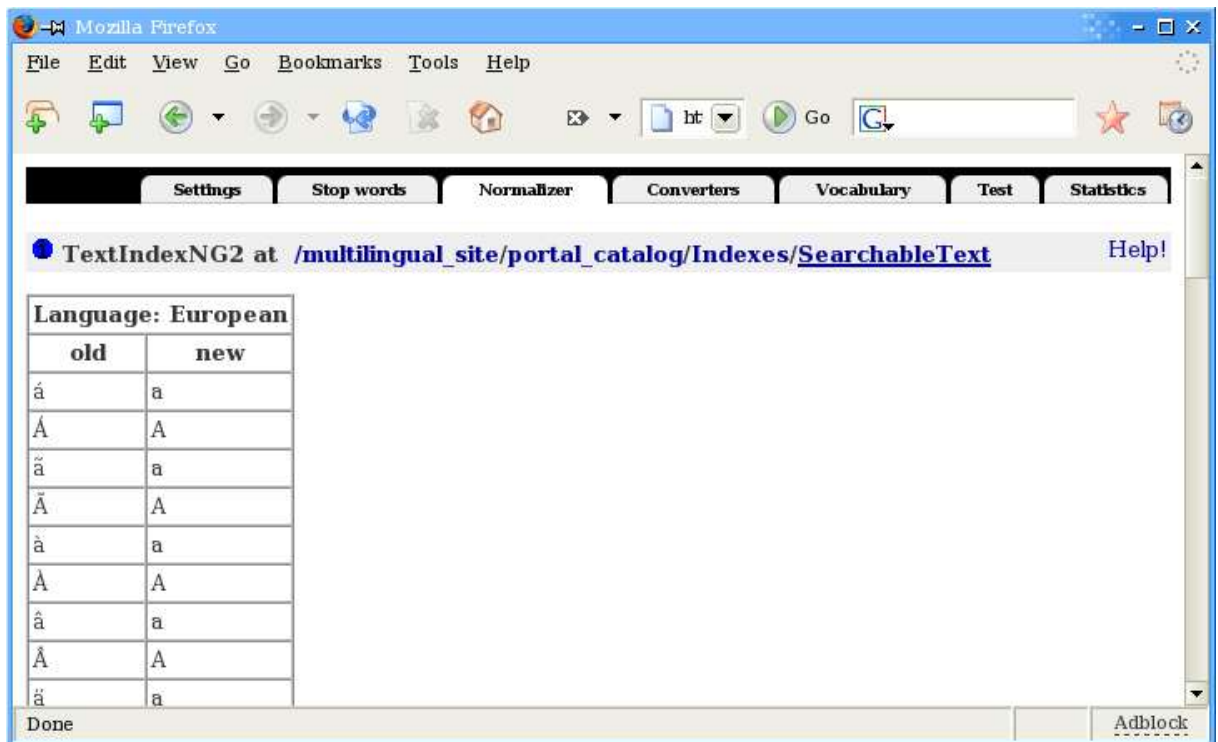
In TextIndexNG we noticed a conflict when creating an index with stemming and at the same time globbing support. Since globbing is more powerful than stemming, we decided to create our indexes without stemming support.

3.3.4.3. Normalizers

Normalization means replacing parts of a string (or special characters) by a more general string, especially umlauts. For example: Ä--> Ae, Ö --> Oe. TextIndexNg has normalizers available for iso-8859-15, French, German and Portuguese. Figure 6 shows in part the European normalizer.

See <http://cvs.sourceforge.net/viewcvs.py/textindexng/TextIndexNG/normalizers>

² <http://snowball.tartarus.org>

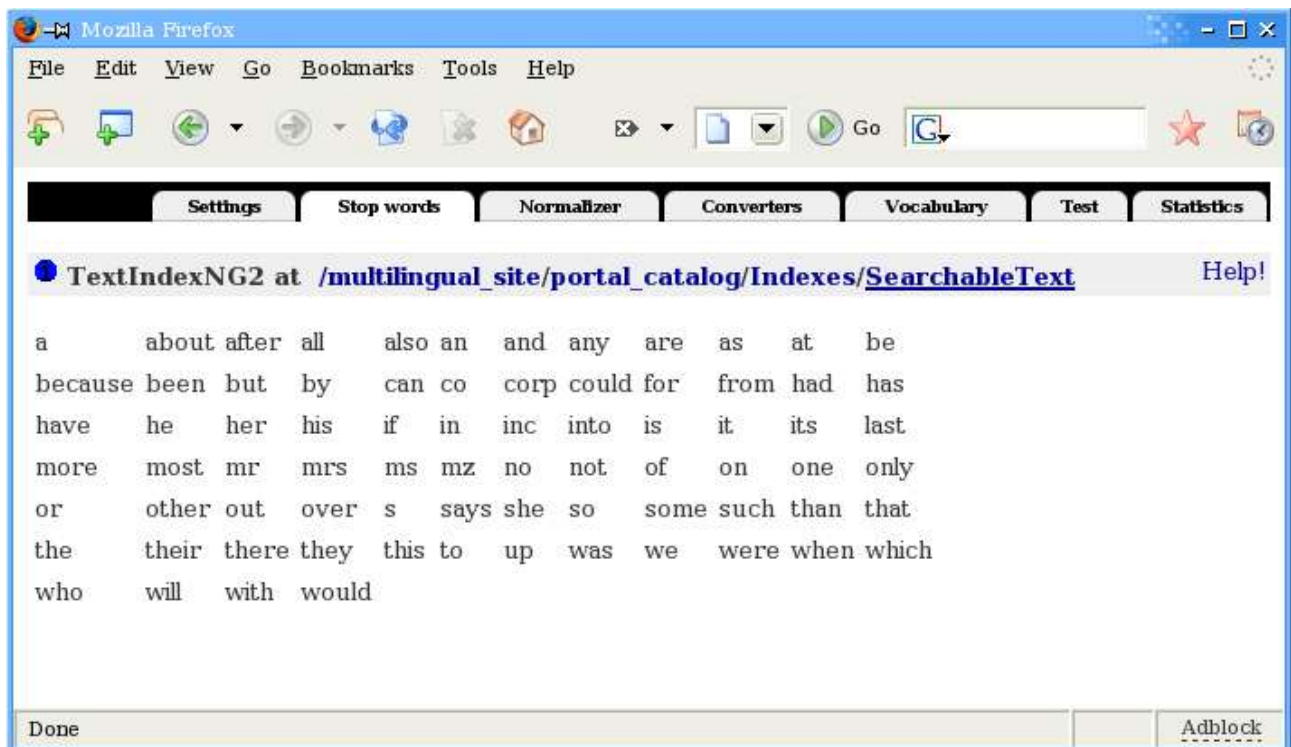


Afbeelding 6 Normalized European characters

3.3.4.4. StopWords

TextIndexNG comes with a set up pre-defined language-dependent stopword files. They can be found in CVS: <http://cvs.sourceforge.net/viewcvs.py/textindexng/TextIndexNG/stopwords/>

Currently eleven languages are supported. And you can very easily define your own stopfiles. Following figure shows a list of English Stop Words.



Afbeelding 7 English set of stop words

3.3.4.5. Similarity Search

Similarity Search means searching for words that are similar or sound similar. The first versions of TextIndexNG provided optional support for the following algorithms: Soundex and Metaphone

<http://en.wikipedia.org/wiki/Soundex>

Soundex, Metaphone, Miracode, and Daitch-Mokotoff Soundex are phonetic algorithms for indexing names by their sound, when pronounced in English. The basic aim is for names with the same pronunciation to be encoded to the same string.

They are necessarily complex algorithms with many rules and exceptions, because English spelling and pronunciation is complicated by historical changes in pronunciation and words borrowed from many languages.

Some soundexing examples :

http://www.archives.gov/research_room/genealogy/census/soundex.html

In the most recent versions of TextIndexNG2, similarity search (soundex, metaphone, doublemetaphone) was dropped and replaced with a more general approach and language independent approach using the **Levenshtein distance**.

According to Wikipedia: http://en.wikipedia.org/wiki/Levenshtein_distance :

“The **Levenshtein distance** or **edit distance** between two strings is the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution. It is named after the Russian scientist Vladimir Levenshtein who invented the term in 1965. It is a generalization of the Hamming distance.”

For example, the Levenshtein distance between “kitten” and “sitting” is 3, since these three edits change one into the other, and there's no way to do it with less than three edits:

kitten
sitten
sittin
sitting

http://en.wikipedia.org/wiki/Hamming_distance :

“The **Hamming distance** is the number of positions in two strings of equal length for which the corresponding elements are different. It measures the number of substitutions required to change one into the other.”

3.3.4.6. PluggableParsers

TextIndexNG comes with a registry for parsers that allow other applications to plugin in their own parsers to transform queries into the internal representation of TextIndexNG. TextIndexNG has the following parsers included:

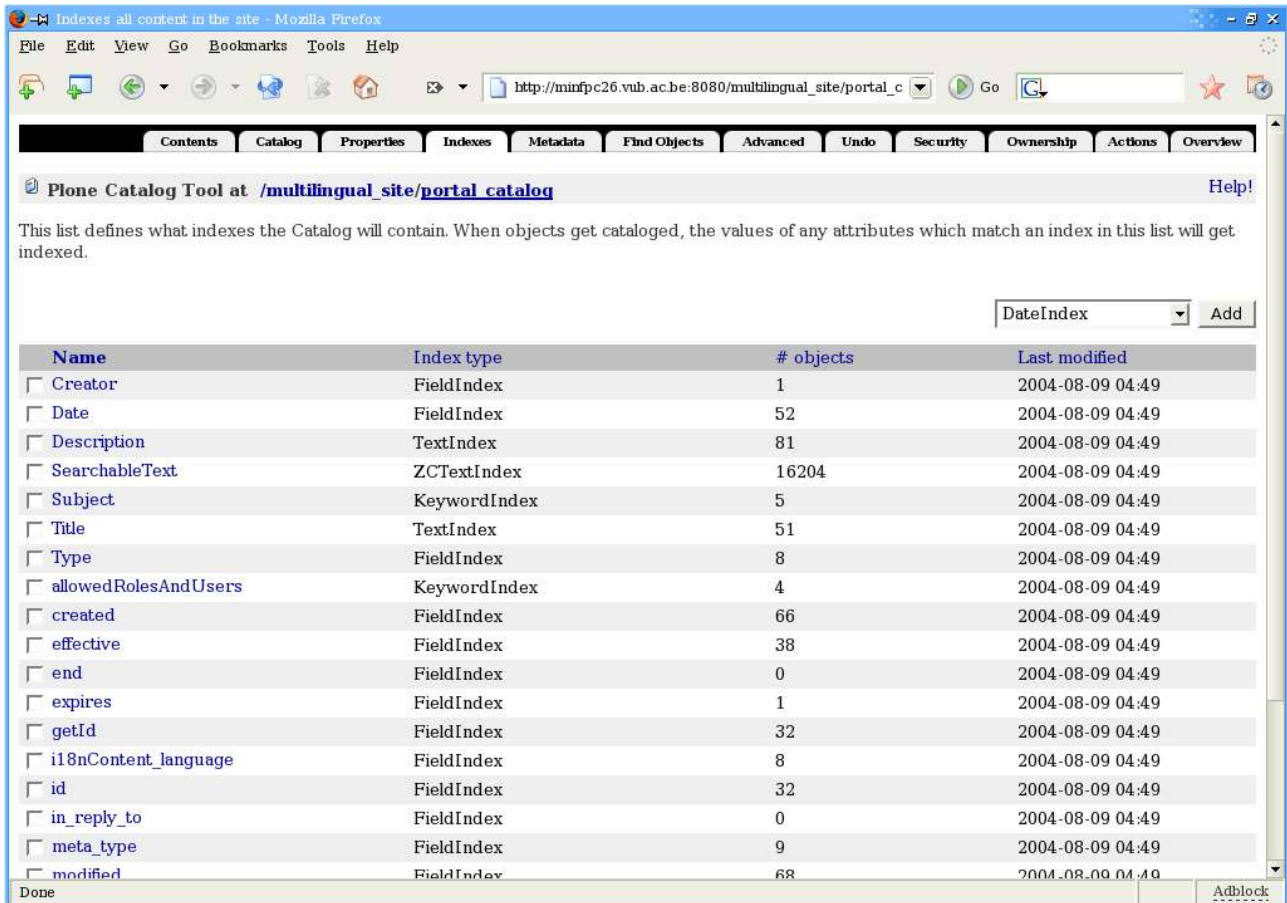
1. **QueryParser** : <http://zope.org/Members/ajung/TextIndexNG/wiki/QueryParser>
2. **DumbQueryParser** : AND, OR, NEAR or QUOTE (phrase search). For most applications DumbQueryParser is sufficient, it's but not as flexible as the more complete QueryParser mentioned above.
<http://zope.org/Members/ajung/TextIndexNG/wiki/DumbQueryParser>
3. **GermanQueryParser** : this supports boolean operators translated into German. It uses "UND", "ODER", "NAHE" and "NICHT" instead of "AND", "OR", "NEAR" and "NOT". <http://zope.org/Members/ajung/TextIndexNG/wiki/GermanQueryParser>

4. Switching from ZCTextIndex to TextIndexNG

A new Plone-website is by default created with ZCTextIndex as default fulltext-indexer. Comparison shows that TextIndexNG is more powerful and feature-rich. Therefore we decide the change the fulltextindex from ZCTextIndex to TextIndexNG. Instruction for doing are clearly outlinde at <http://zope.org/Members/ajung/TextIndexNG/wiki/InstallationInstructions>

After installation and solving the necessary dependencies that fall outside the scope of this paper, we go to the index-tab of the portal_catalog, shown in figure 15.

“SearchableText” is the only ZCTextIndex present. (Since there is only one fulltext-index)This SearchableText contains 16204 objects.



Afbeelding 8 SearchableText is a ZCTextIndex

Now to replace it with TextIndexNG

1. Select the SearchableText-index and remove it (delete-button is visible in figure 8)
2. Then select TextIndexNG from the dropdown-box
3. Now you get a fill-in form, an “id” is asked. Fill in id = “SearchableText”. In Plone you must add TextIndexNG as a SearchableText- index if you want it to give results from the search box in the user-interface
4. Change the default encoding from ISO-8859-10 to UTF-8
5. Select the English Stemmer (Porter)
6. We set Near-search distance to “10”
7. Enable left truncation and autoexpansion,
8. Select an English stopwords File, European Normalization,
9. Enable Document converters

Add TextIndexNG2

TextIndexNG2 breaks text up into individual words, and are often referred to as full-text indexes.

Id

Indexed attributes *attr1,attr2,... or leave empty*

Default encoding

Storage

Stemmer

Splitter: max length of splitted words

Splitter: index single characters

Splitter: casefolding

Splitter: allowed characters inside a word (e.g. 'C++', 'python-22.lib')

Near search distance

Left truncation

Query term autoexpansion

Stopwords File

Normalization

Document converters (PDF, Word etc.)

Type

Done Adblock

Afbeelding 9 Configuration-form when adding a new TextIndexNG-index

When this form is filled in and submitted, you see the same image as in figure 8, except with SearchableText is now an TextIndexNG with zero objects.

Again, select the SearchableText and click and "Reindex"-button, to reindex the entire website with the TextIndexNG-index.

We will play with these settings and therefore need to redo all these steps many times, in order to test and compare these indexers with each other, and to find an optimal setting for our multilingualwebsite. Figure 9 shows the fill form of a new TextIndexNG.

After indexing the entire site, we can check some statistics. Figure shows we currently have 66 iondexed documents totalling 21 996 words in the vocabulary

Settings Stop words Normalizer Converters Vocabulary Test Statistics

TextIndexNG2 at [/multilingual_site/portal_catalog/Indexes/SearchableText](#) [Help!](#)

Statistics

indexed documents 66

words in vocabulary 21996

Afbeelding 10 Indexing Statistics

Settings	Stop words	Normalizer	Converters	Vocabulary	Test	Statistics
TextIndexNG2 at /multilingual_site/portal_catalog/Indexes/SearchableText Help!						
Show words in vocabulary starting with <input type="text" value="a"/>		977 matches		a b c d e f g h i j k l m n o p q r s t u v w x y z		
<input type="button" value="Submit Query"/>						
a	a-	a-rationeel	a-rationele	a1692.g.akamai.net		
aachen	aan	aanbrengen	aandacht	aanduiden		
aangedaan	aangeduid	aangehangen	aangelegd	aangelegenheden		
aangename	aangenomen	aangetoond	aangetrokken	aangewezen		
aangezet	aangezien	aanhang	aanhangers	aanhangt		
aanhechting	aanhoren	aanhoudende	aanklager	aanleiding		
aanliggende	aanlopen	aanneemt	aanmemelijk	aannemen		
aanpak	aanpast	aanraken	aanreikt	aanschouwing		
aansloot	aansporingen	aanspraak	aantal	aantallen		
aantasten	aantonen	aantoot	aantreffen	aantreft		
aanvaard	aanvaardbaar	aanvaarde	aanvaardden	aanvaardt		
aanvankelijk	aanwenden	aanwezig	aanzet	aard		
aardbeving	aardbevingen	aarde	aardse	aarden		
ab	abandoned	abbey	abci	abdanken		
abdicate	abdiquer	aber	abflaute	abgebaut		
abgefangen	abgelehnt	abgeleitete	abgelost	abgeordneten		
abgeordnetenammer	abgesehen	abgesichert	abkommen	able		

Afbeelding 11 Resulting vocabulary after indexing

Figure 11 shows the vocabulary starting with letter “a”. Clicking on any word, will lead to a list of documents containing that very word. For example, figure 12 shows the list of documents containing the term “Japan”.

TextIndexNG2 at /multilingual_site/portal_catalog/Indexes/SearchableText
All documents containing the word japan :
/multilingual_site/articles/germany/en
/multilingual_site/articles/germany/nl
/multilingual_site/articles/japan/ar
/multilingual_site/articles/japan/de
/multilingual_site/articles/japan/en
/multilingual_site/articles/japan/ja
/multilingual_site/articles/japan/nl
/multilingual_site/articles/russia/de
/multilingual_site/articles/russia/en
/multilingual_site/articles/russia/ja

Afbeelding 12 List of documents containing the term "Japan"

5. Empirical Comparison of TextIndex and TextIndexNG2

In order to easily compare a website driven with the ZCTextIndex back-end and the TextIndexNG backend, we make a copy of the Plonesite and change the fulltext-index in one of them in order to be able to access both sites simultaneously.

In our instance, we even have two zope-servers running on the same server. The URLs are respectively:

- http://minfpc26.vub.ac.be:9080/multilingual_site (ZCTextIndex Driven)
- http://minfpc26.vub.ac.be:8080/multilingual_site (TextIndexNG Driven)

Test 1 : Stop Words Test

We select a few of these keywords and search them on both our sites

Following two figures shows the search result for "the". The succesful test shows that with ZCTextIndex many results are found, but in with TextIndexNG the stop words-feature succesfully succesfully result in no results, as wanted.

Zoekresultaten

Niet gevonden waar u naar zoekt? Probeer de [uitgebreide zoekresultaten](#) voor preciezere zoekopties.

18 items die aan de zoekterm(en) voldoen. [RSS](#)

 [Japan](#) [100%] van admin, 2004-08-04 01:36 PM

 [Germany](#) [99%] van admin, 2004-08-04 04:30 AM

 [Belgium](#) [99%] van admin, 2004-08-04 06:26 AM

Wikipedia Artikel on Belgium, this text is licensed under the GNU Free Documentation License

 [Belgium](#) [99%] van admin, 2004-07-15 06:53 AM

 [Russia](#) [99%] van admin, 2004-08-04 01:42 PM

 [France](#) [99%] van admin, 2004-08-04 01:29 PM

 [Paper](#) [99%] van admin, 2004-08-02 11:08 PM

 [My multilingual little project](#) [96%] van admin, 2004-08-02 11:14 PM

This website is presented as a use-case of the Plone Content Management System to quickly develop multilingual websites.

 [Mijn klein meertalig projectje](#) [94%] van admin, 2004-08-02 11:18 PM

Deze website toont een voorbeeld van hie meertalige website ontwikkeld kunnen worden in het Plone Content Management System.

 [Japan](#) [85%] van admin, 2004-08-04 01:35 PM

 [Frankrijk](#) [78%] van admin, 2004-08-04 01:30 PM

 [Duitsland](#) [69%] van admin, 2004-08-04 06:28 AM

Afbeelding 13 Searching for "the" with TextIndex





Test 2 : Test Normalizer

Searching for "Bevolkerung" or "Bevölkerung" yields exact the same results with ZCTextindex and TextIndexNG, since but support normalisation. We don't have any text on our multilingual site to discriminate between the two indexers. But since TextIndexNG is more modular, and more easier to add different languages to , more scalability to more languages is expected.

Search results

Did you not find what you were looking for? Try the [Advanced Search](#) for more precise search options.

4 items matching your criteria. [RSS](#)

-  [Deutschland](#) [100%] by admin, 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Deutschland>
-  [Frankreich](#) [75%] by admin, 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Frankreich>
-  [Russland](#) [73%] by admin, 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Russland>
-  [Japan](#) [71%] by admin, 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Japan>

Afbeelding 15 Search-results for "bevolkerung"

Test 3: Stemming Test

We first need to find some words
we cut and and paste veral of our english docmuents










<http://snowball.tartarus.org/demo.php>

As you can read you can use a query with wildcards if stemming is enabled (Stemming and wildcard searches don't make sense). This is a feature and not a bug. Search for non-latin terms. Since this website is truely multilingual and we changed the default encoding of our indexer to UTF-8 we can search terms in other alphabets than just latin. Figure shows search-results for the russian term of "Russia"

Résultats de la recherche

Avez-vous trouvé ce que vous cherchiez ? Essayez la [recherche avancée](#) pour des critères de recherche plus précis.

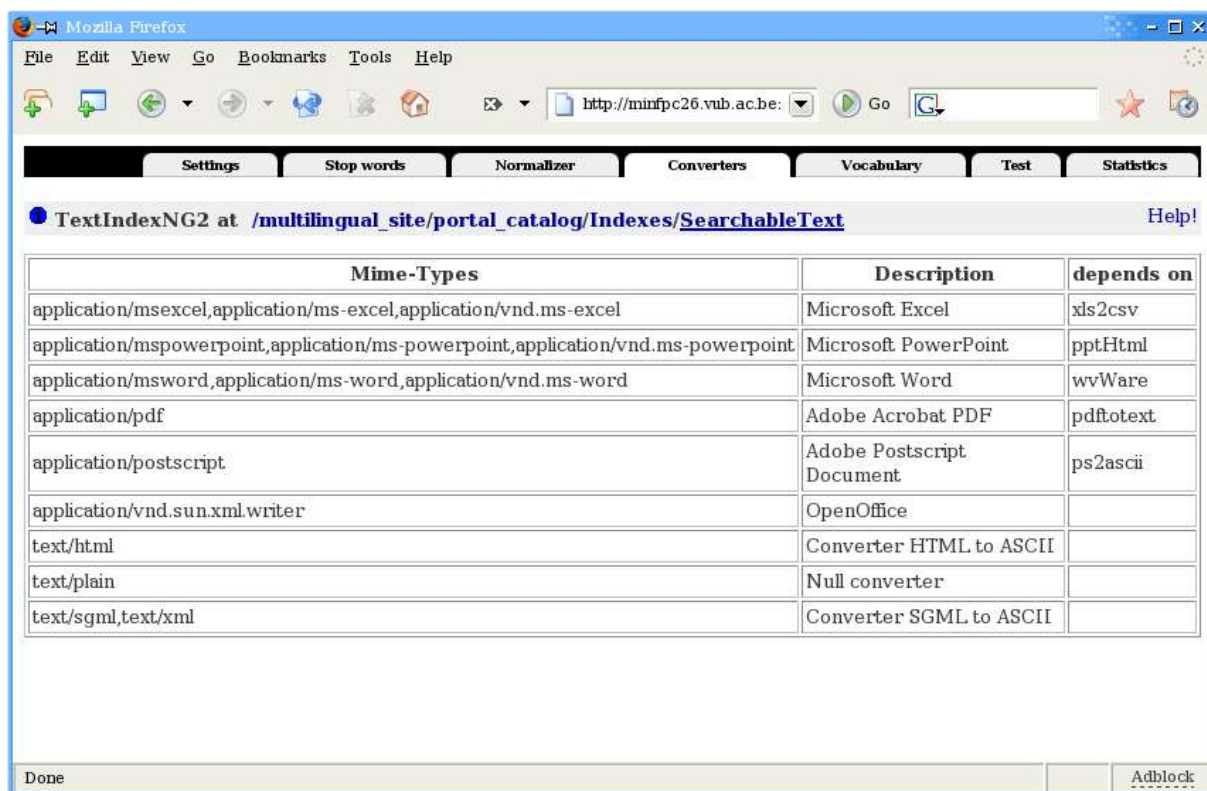
Il y a 9 éléments qui répondent à vos critères. [RSS](#)

-  [Россия](#) [100%] par admin, le 2004-08-04 01:42 PM
-  [Япония](#) [30%] par admin, le 2004-08-06 03:35 PM
Если вы заметили ошибку в этой статье или можете дополнить её — нажмите кнопку ...
-  [Франция](#) [30%] par admin, le 2004-08-04 01:29 PM
-  [Бельгия](#) [30%] par admin, le 2004-08-04 06:27 AM
-  [Бельгия](#) [30%] par admin, le 2004-07-15 06:55 AM
-  [Russland](#) [30%] par admin, le 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Russland>
-  [30\] رويسيا](#) [30%] par admin, le 2004-08-04 01:44 PM
-  [Германия](#) [30%] par admin, le 2004-08-04 06:28 AM
-  [日本](#) [30%] par admin, le 2004-08-04 01:40 PM

Test 4 : Convertors

Supported file-formats are: PDF, Postscript, XLS (Excell), .SXI (OpenOffice/Oasis), DOC, HTML, SGML. To test this, we need some new documents in these formats.

For testing purposes, we upload the powerpoint-file of Nicolas Van Vosselen's Symbolic Logic lectures, and a few pdf's with chapters of the course of Prof Braeckman's Historic Overview of



Afbeelding 16 Available convertors and supported file-formats

Philosophy, to be found at <http://www.flwi.ugent.be/philosophy/histowb>.

After uploading new documents, we query for some terms only present in these documents. With TextIndex we get no results as expected, since it can't index these formats. But with TextIndexNG2 we get succesful results, shown in figure 11, for a search on "Logic", present in the Powerpoint-file as well as the pdf's.

Search results

Did you not find what you were looking for? Try the [Advanced Search](#) for more precise search options.

3 items matching your criteria. [RSS](#)

[symbolic_logic.ppt](#) [100%] by admin, 2004-08-09 04:21 PM

[hoofdstuk 2.pdf](#) [54%] by admin, 2004-08-09 04:04 PM

[hoofdstuk 1.pdf](#) [50%] by admin, 2004-08-09 04:04 PM

Afbeelding 17 Search for "Logic" in binary documents









Test 5: Globbing

We can search for "Symbolic Logic", but if we aren't sure on something, we can use wild-cards *. With ZCTindex we get no results, since globbing is not supported. Figure 11 shows the succesful result of searching for "Symb*" with TextIndexNG.

Search results

Did you not find what you were looking for? Try the [Advanced Search](#) for more precise search options.

8 items matching your criteria. [RSS](#)

-  [symbolic_logic.ppt](#) [100%] by admin, 2004-08-09 04:21 PM
-  [hoofdstuk_2.pdf](#) [55%] by admin, 2004-08-09 04:04 PM
-  [Japon](#) [46%] by admin, 2004-08-04 01:35 PM
-  [hoofdstuk_1.pdf](#) [38%] by admin, 2004-08-09 04:04 PM
-  [Duitsland](#) [34%] by admin, 2004-08-04 06:28 AM
-  [Allemagne](#) [23%] by admin, 2004-08-04 06:28 AM
-  [Japan](#) [21%] by admin, 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Japan>
-  [Japan](#) [21%] by admin, 2004-08-04 01:36 PM

Afbeelding 18 Search for "Symb*"

Test 6 : Auto-expansion

Auto-expansion expands a query term "foo" to "foo*", if there was no hit for "foo". So this feature extends the globbing-functionality further.





In ZCTextIndex, searching for "duit" yields no results because there is no word "duit" and auto-expansion is not supported.

But in TextIndexNG, when no results are found, then automatically the globbing-operator "*" is added to term. Figure 14 shows a succesful auto-expanded search for "duit*". So using an asterisk * to initiate a globbing-search is not even necessary.

Search results

Did you not find what you were looking for? Try the [Advanced Search](#) for more precise search options.

4 items matching your criteria. [RSS](#)

-  [Duitsland](#) [100%] by admin, 2004-08-04 06:28 AM
-  [België](#) [54%] by admin, 2004-08-06 03:37 PM
-  [België](#) [54%] by admin, 2004-07-25 04:02 AM
-  [Frankrijk](#) [5%] by admin, 2004-08-04 01:30 PM

Afbeelding 19 Auto-expanded search-results for "duit" in TextIndexNG









Test 7 : Left-truncation

Globbing-support can be extended even further. TextIndexNG can be configured (at creation-time of the index) to support left-truncation. This means you can search for *suffixes. Figure 1 shows the search-result for "*land". Searching based on ZCTextIndex does not yield any results, since left-truncation there is not supported.

Search results

Did you not find what you were looking for? Try the [Advanced Search](#) for more precise search options.

24 items matching your criteria. [RSS](#)

-  [Russland](#) [100%] by admin, 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Russland>
-  [Deutschland](#) [81%] by admin, 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Deutschland>
-  [Duitsland](#) [52%] by admin, 2004-08-04 06:28 AM
-  [Japan](#) [43%] by admin, 2004-08-04 01:36 PM
-  [Russia](#) [32%] by admin, 2004-08-04 01:42 PM
-  [Germany](#) [31%] by admin, 2004-08-04 04:30 AM
-  [France](#) [25%] by admin, 2004-08-04 01:29 PM
-  [Belgium](#) [21%] by admin, 2004-08-04 06:26 AM
Wikipedia Artikel on Belgium, this text is licensed under the GNU Free Documentation License

Afbeelding 20 Search for "*land" with TextIndexNG

Test 8 : Phrase Search

Phrase-search conflicts with the use of a stopwords file. This was to be expected because stopwords are deleted in the search-string is scanned for. So probabilities become very small to actually get any result.

To use phrase search, you shouldn't make use of stopwords. Since ordinary users make frequently and intuitively use of phrase search, the stopwords-feature should not be used for normal websites.

Test 9 : Similarity Search

Despite much effort, similarity-search based on the Levenshtein-distance doesn't seem to work in the current version of TextIndexNG. Some bug must have crept in.

Test 10 : Range Search

TextIndexNG supports range searches like "Fi..Foo"

For example, a query for "Belgium" can be put in range like "Ba..Bo". Figure 15 shows the search-result for this query with "Ba..Bo". With ZCTextIndex we get no results, as expected. With TextIndexNG we get as much as 38 results, shown below.

Search results

Did you not find what you were looking for? Try the [Advanced Search](#) for more precise search options.

38 items matching your criteria. [RSS](#)

-  [hoofdstuk 1.pdf](#) [100%] by admin, 2004-08-09 04:04 PM
-  [hoofdstuk 2.pdf](#) [99%] by admin, 2004-08-09 04:04 PM
-  [Deutschland](#) [42%] by admin, 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Deutschland>
-  [Russland](#) [21%] by admin, 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Russland>
-  [Japan](#) [16%] by admin, 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Japan>
-  [Belgium](#) [12%] by admin, 2004-08-04 06:26 AM
Wikipedia Artikel on Belgium, this text is licensed under the GNU Free Documentation License
-  [Belgium](#) [12%] by admin, 2004-07-15 06:53 AM

Afbeelding 21 Succesful Range Search "Ba..Bo"

Test 12 : Near search

Both ZCTextIndex and ZCTexindexNG support the “near”-operator in boolean search, but in TextIndexNG the word-distance of what is considered “near” can be fine-tuned. Following figures shows a search-result for respectively “France country” and “France near country”.

The first query yields 5 matches, and the second only 2. Checking these results shows that in the two remaining result “France” is effectively near the word “country”.

Search results

Did you not find what you were looking for? Try the [Advanced Search](#) for more precise search options.

5 items matching your criteria. [RSS](#)

 [Belgium](#) [100%] by admin, 2004-08-04 06:26 AM

Wikipedia Artikel on Belgium, this text is licensed under the GNU Free Documentation License

 [Germany](#) [96%] by admin, 2004-08-04 04:30 AM

 [France](#) [80%] by admin, 2004-08-04 01:29 PM

 [Russia](#) [63%] by admin, 2004-08-04 01:42 PM

 [Japan](#) [63%] by admin, 2004-08-04 01:36 PM

Afbeelding 22 Search for "France country"

Search results

Did you not find what you were looking for? Try the [Advanced Search](#) for more precise search options.

2 items matching your criteria. [RSS](#)

 [Belgium](#) [100%] by admin, 2004-08-04 06:26 AM

Wikipedia Artikel on Belgium, this text is licensed under the GNU Free Documentation License

 [France](#) [80%] by admin, 2004-08-04 01:29 PM

Afbeelding 23 Search for "France near country", with a word-distance of 15

This near-search was performed with a configured word-distance of 15. This means there can be as much as 15 words in between the two inputted words. The distance for the first result, Belgium, is 9. While for the second result, France, the distance is only 2.

We can fine-tune this distance via the TextIndexNG-settings. This settings can be changed without reindexing the entire site. Figure 21 shows this distance set to 5.

Settings	Stop words	Normalizer	Converters	Vocabulary	Test	Statistics
----------	------------	------------	------------	------------	------	------------

TextIndexNG2 at /multilingual_site/portal_catalog/Indexes/SearchableText [Help!](#)

1

Indexed attributes	SearchableText,data
Default encoding	UTF-8
Storage	StandardStorage
Stemmer	disabled
Splitter: casefolding	enabled
Splitter: index single characters	enabled
Splitter: max. length of splitted words	64
Splitter: separator characters	.+ _@
Default query parser	PyQueryParser
Autoexpansion	enabled (Limit: 4)
Stopwords	None
Normalizer	European
Use converters	enabled
Near distance	<input type="text" value="5"/>
Left truncation	enabled
Debug mode	<input type="text" value="off"/>

Afbeelding 24 Settings of TextIndexNG with near-distance set to 5
 We now issue the same near-query again. As hoped for, we now get only 1 result, shown in figure 22.

Search results

Did you not find what you were looking for? Try the [Advanced Search](#) for more precise search options.

1 items matching your criteria. [RSS](#)

 [France](#) [100%] by admin, 2004-08-04 01:29 PM

Afbeelding 25 Search for "France near country", with a word-distance of 5

Test 13 : Multilingual Searches (UTF-8)

Searching with Russian, Arabic and Japanese terms seems to have no problems at all. Also globbing works seems to work flawlessly. Following figures shows the result of some search-queries in different character-sets.

Результаты поиска


Вы не нашли то, что искали? Попробуйте [расширенный поиск](#) для более тонких настроек поиска.

элементов соответствующих Вашему критерию [RSS](#)

 [Франция](#) [100%] admin, 2004-08-04 01:29 PM

 [Бельгия](#) [16%] admin, 2004-08-04 06:27 AM

 [Германия](#) [10%] admin, 2004-08-04 06:28 AM

 [Россия](#) [3%] admin, 2004-08-04 01:42 PM

Afbeelding 26 A globbed search-result for a truncated "France" in Russian

ش حبالا ج نأند

تقد ر شكأ ش حبالا ت ارايذ داجيلا جديتم ش حديم اندتسأل واحد هنع ش حبتت نك امدجت م اذا

RSS مديحتيت مفا مع م ق باطتتي نلا ر صانعلا 2

01:35 04-08-2004 [%100] بيان PM admin

01:30 04-08-2004 [%100] العربية PM admin

Afbeelding 27 Search-results for "France" in Arabic

My Multilingual Little Project

1830 検索

ホーム ニュース メンバー

admin :マイフォルダ :個人設定 :元に戻す :phoneの設定 :ログアウト

現在の場所: ホーム » 検索結果

ナビゲーション

- ホーム
- Members
- 記事
- Papers
- Searching and Indexing

最近公開のアイテム

前回ログイン後に公開されたアイテムはありません。

もっと...

zoeken

検索結果

お探しのコンテンツが見つからない場合、[アドバンスド検索](#)にてさらに詳細な検索オプションをお試しください。

8 個のアイテムが検索条件に該当しました。RSS

- Belgium [100%] 作成者: admin 最終変更日: 2004-08-04 06:26 AM
Wikipedia Artikel on Belgium, this text is licensed under the GNU Free Documentation License
- België [91%] 作成者: admin 最終変更日: 2004-08-06 03:37 PM
- Belgique [80%] 作成者: admin 最終変更日: 2004-08-04 06:27 AM
- フランス [64%] 作成者: admin 最終変更日: 2004-08-04 01:30 PM
- ベルギー [64%] 作成者: admin 最終変更日: 2004-08-04 06:27 AM
- Opnamme [38%] 作成者: admin 最終変更日: 2004-08-04 01:29 PM
- Belgien [38%] 作成者: admin 最終変更日: 2004-08-04 06:27 AM
- Belgien [38%] 作成者: admin 最終変更日: 2004-08-05 12:43 PM
Der Inhalt dieser Seite steht unter der GNU Free Documentation License. <http://de.wikipedia.org/wiki/Belgien>

Afbeelding 28 Mixed character-set in the search-result for the year "1830" (Japanese language selected)

6. Our Optimal Configuration

To conclude this comparison between ZCTextIndex and TextIndexNG, and our playing around with various configurations of TextIndexNG, we can reach some conclusions about an optimal configuration for our multilingual website. This optimal configuration is as follows:

1. A fulltext-index based on TextIndexNG
2. with an UTF-8 encoding for international character-sets.
3. Casefolding enabled (ignore capital letters)
4. Disable the stopwords (sabotages searching for phrases)
5. Disable stemming (globbing is more useful)
6. We allows for single character to be indexed (since this is a small site. For bigger sites, this should be disabled to keep the index significantly smaller)
7. Auto-expansion and left-truncation enabled
8. The convertors enabled (certainly if you upload pdf's and Office-documents)

9. A near-distance of 15

7. Thesaurus

TextIndexNG2 currently does not support thesauruses. However, in August 2003, Andreas Jung says the following on the Zope-CMF mailing-list in reply to a question about thesaurus-support:

“I thought about adding thesaurus support to TextIndexNG but I dropped this idea because there was no demand for this feature. However it would not be too hard to add such a feature.”³

Wordnet is a very well-known thesaurus for the English Language. For European Languages like Dutch, Italian, Spanish, German, French, Czech and Estonian, the **EuroWordNet** has done a similar effort. The cooperative framework of EuroWordNet is now continued through the **Global WordNet Association**. This is a free and public association that builds on EuroWordNet and Princeton WordNet. The aim is to stimulate further building of wordnets, further standardization and interlinking and the development of tools, dissemination of information.

To jump-start thesaurus-support in TextIndexNG and Zope that is written in the Python Programming Language, one could start from **PyWordnet**, which is a Python interface to the WordNet database.

1. Wordnet : <http://www.cogsci.princeton.edu/~wn>
2. EuroWordNet : <http://www.ilc.uva.nl/EuroWordNet>
3. GlobalWordNet : <http://www.globalwordnet.org>
4. PyWordNet : <http://osteele.com/projects/pywordnet/>

³ <http://mail.zope.org/pipermail/zope-cmf/2003-August/019234.html>

8. Integrating Office Documents in Plone CMS

As the course this paper is submitted for, is about content and document management, we briefly touch upon another interesting and relevant product.

This paper is written in OpenOffice.org (www.openoffice.org), a freely downloadable office-suite similar and largely compatible with Microsoft Office. A PDF-file was generated simply by exporting it in OpenOffice.org to PDF.

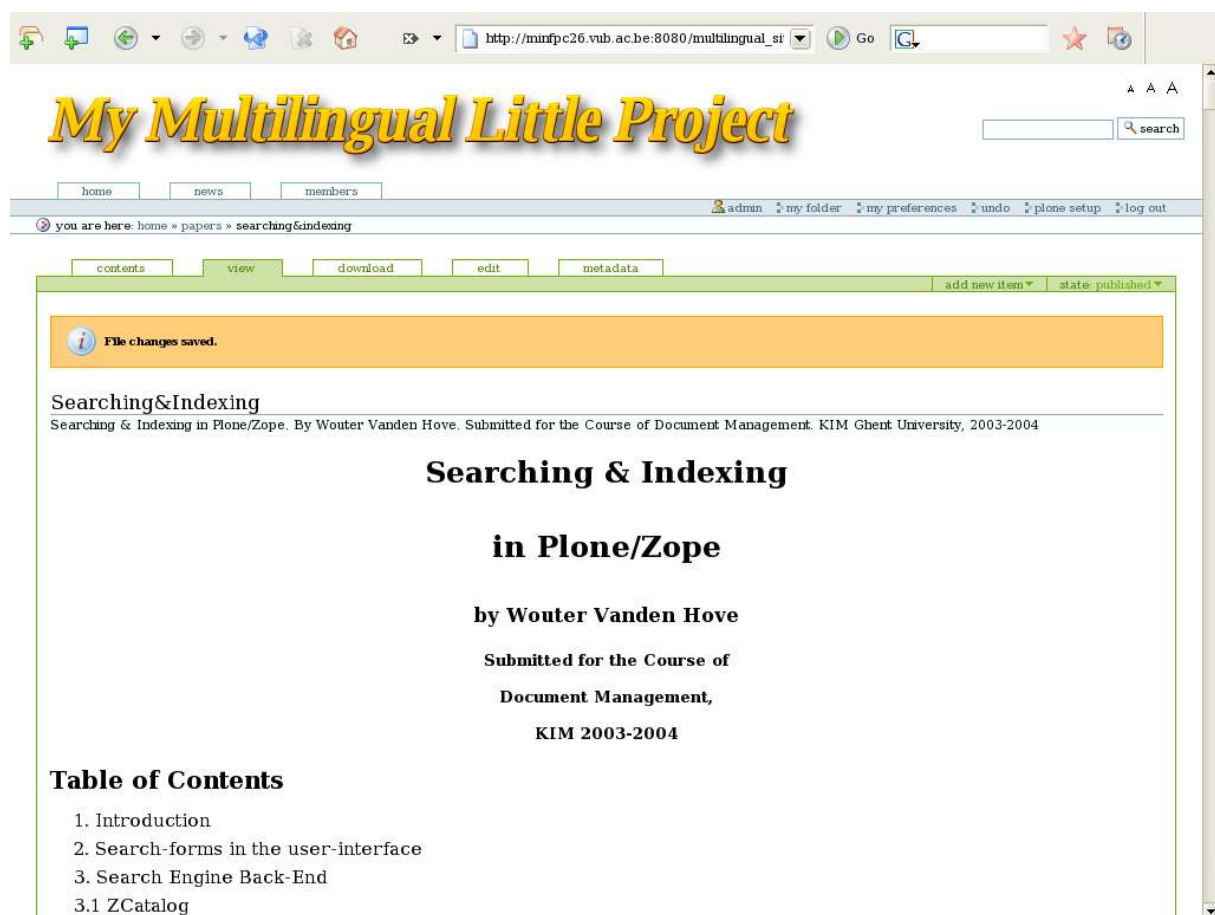
CMFOODocument⁴ (short for CMF OpenOffice Document) is a Zope-Product that integrates OpenOffice.org Documents including images, and tables into a CMF or Plone powered Website.

The native XML-based fileformat of OpenOffice is destined to become an open standard⁵ under the auspices of OASIS⁶, the Organization for the Advancement of Structured Information Standards), a not-for-profit, international consortium that drives the development, convergence, and adoption of e-business standards.

CMFOODocument provides for a new type of document that can be uploaded by end-users of the plone-site. When uploading an OpenOffice document, it is stored on the server as the original OpenOffice document, without any conversion. But when accessing this file via an url in a webbrowser, it is converted on the fly to HTML via an XLST-transformation. It is just as easy for download and editing by clicking on the download-tab, visible in the screenshot below.

Following figure shows this very paper uploaded as an OpenOffice-document

http://minfpc26.vub.ac.be:8080/multilingual_site/papers/searching_plone.sxw/file_view



Afbeelding 29 This paper uploaded as an OpenOffice-Document

CMFOODocument is a very elegant way to integrate Office documents into a web-based content management system.

⁴ <http://www.zope.org/Members/longsleep/CMFOODocument>

⁵ http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office

⁶ <http://www.oasis-open.org>

9. Other Open Source Indexing Software

In this paper we explored the search & indexing system of the Zope Application Server. There are however many other open source indexing-software projects. Below is a short non-exhaustive list of such projects:

- **Xapian** : <http://www.xapian.org/>

Xapian is an Open Source Probabilistic Information Retrieval library, released under the GPL. It's written in C++, and bindings are under development to allow use from other languages (Perl, Python, and PHP are working; Java will be available shortly). Xapian is designed to be a highly adaptable toolkit to allow developers to easily add advanced indexing and search facilities to their own applications.

- **Jakarta Lucene** : <http://jakarta.apache.org/lucene/docs/index.html>

Jakarta Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

- **ht://Dig** : <http://www.htdig.org>

The ht://Dig system is a complete world wide web indexing and searching system for a domain or intranet.

- **ZAP/Zebra** : <http://www.indexdata.dk/zap> , <http://www.indexdata.dk/zebra>

Zebra is a high-performance, general-purpose structured text indexing and retrieval engine. It reads structured records in a variety of input formats (eg. email, XML, MARC) and allows access to them through exact boolean search expressions and relevance-ranked free-text queries. Zebra supports large databases (more than ten gigabytes of data, tens of millions of records). It supports incremental, safe database updates on live systems. You can access data stored in Zebra using a variety of Index Data tools (eg. [YAZ](#) and [PHP/YAZ](#)) as well as commercial and freeware Z39.50 clients and toolkits.

- **Z39.50 Resource Page** : <http://www.niso.org/z39.50/z3950.html>

ANSI/NISO Z39.50 defines a standard way for two computers to communicate for the purpose of information retrieval. Z39.50 makes it easier to use large information databases by standardizing the procedures and features for searching and retrieving information.

- **E-SWISH** : <http://swish-e.org>

SWISH-E is a fast, powerful, flexible, free, and easy to use system for indexing collections of Web pages or other files.

10. Conclusion

In this paper we studied in detail the search-engine underlying the Plone content management system. We showed how an entire website can be indexed with various specialized indexers, and we showed how these can affect search-results by comparing the default indexing system with a more advanced one.

We have shown how various file-formats like *.pdf, *.ppt and *.doc can be indexed succesfully. This succesful result has prompted me to replace the default indexer ZCTextIndex with the more advanced TextIndexNG on all of my deployed Plone2 sites, like www.skepp.be, www.etiennevermeersch.be.

For our thesis-work on the KIM Student Center (<http://kim.opencursus.org>) it should be noted that this site runs on Plone1 where TextIndexNG cannot be used because of dependencies requiring Plone2. Therefore, an upgrade to Plone 2 is necessary first.

Unfortunately, thesaurus-support is for the moment not available. But this is only a matter of time, since the Plone CMS is rapidly maturing, corporate adopting is growing and the core developer of TextIndexNG has put on top of his todo-list.

An interesting exercise we leave for the reader, is to compare our indexing-system with various other websites. Do search-engines of other websites support globbing? Can they index their own pdf's or docs? Without performing a valid statistical analysis of 1000 other sites and ten of different CMS's, I'm quite confident our system performs above average.

A last remark is that Plone comes with a good pre-configured indexing system, but this is very modular and build upon various modules. It is possible to create the whole indexing from scratch and tailor it completely to your own needs. We refer to chapter 17 of the Zope Book [4] for a complete tutorial on this.

11. References

1. What is RSS? By Mark Pilgrim. December 18, 2002
<http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html>
2. RSS 2.0 Specification, <http://blogs.law.harvard.edu/tech/rss>
3. Managing Gigabytes: Compressing and Indexing Documents and Images. Second Edition, May 1999, published by Morgan Kaufmann Publishing, San Francisco, ISBN 1-55860-570-3.
<http://www.cs.mu.oz.au/mg>
4. The Zope book, Chapter 17: Searching and Categorizing Content
http://zope.org/Documentation/Books/ZopeBook/2_6Edition/SearchingZCatalog.stx
5. How to Index Anything. By Josh Rabinowitz. July 01, 2003
<http://www.linuxjournal.com/article.php?sid=6652>
6. Comparing Open Source Indexers. By Eric Lease Morgan. May 29, 2001
<http://www.infomotions.com/musings/opensource-indexers>
7. Open Source Search Engines : <http://www.searchtools.com/tools/tools-opensource.html>
8. Okapi at {TREC}Text. By Stephen E. Robertson and Steve Walker and Micheline Hancock-Beaulieu and Aarron Gull and Marianna Lau. {REtrieval} Conference, pages 21-30. 1992.
<http://citeseer.ist.psu.edu/article/robertson96okapi.html>