



***Muscat FX 1.4
Manual***



Muscat Limited has endeavoured to provide trademark information about all companies and products mentioned in this manual.

Adobe and Acrobat are trademarks of Adobe Systems Incorporated.

Netscape and Netscape Navigator are trademarks of Netscape Communications Corporation.

Windows, Windows NT, Excel, Word and PowerPoint are trademarks of Microsoft Corporation.

Muscat Limited

The Westbrook Centre

Milton Road

Cambridge

CB4 1YG

Phone: +44 1223 715000

Fax: +44 1223 715001

e-mail: information@muscat.com

Web: <http://www.muscat.com>



Contents

| | |
|--|----|
| Muscat FX 1.4 Manual | 1 |
| Introduction | 5 |
| About this guide | 6 |
| About Muscat FX | 8 |
| Section 1 | |
| Getting started | 9 |
| Installing the software | 11 |
| FX system architecture | 12 |
| Before starting the installation | 13 |
| Installing under Unix | 15 |
| Installing under Windows NT | 18 |
| What next? | 23 |
| Basic configuration | 24 |
| Setting up mappings | 25 |
| Section 2 | |
| Basic use | 29 |
| Creating an index | 30 |
| Setting up the look and feel | 31 |
| Configuring an index | 35 |
| Configuring a single-site index | 35 |
| Advanced configuration options | 45 |
| Building an index | 47 |
| Updating and rebuilding indexes | 48 |
| Using the index browser | 49 |
| Browser information | 50 |
| Indexing controls | 52 |



| | |
|---|----|
| Search controls | 53 |
| Searching an index | 54 |
| Displaying a search form | 55 |
| Submitting a query | 56 |
| Section 3 | |
| Advanced use | 59 |
| Running multiple indexes | 61 |
| Customising FX | 63 |
| Search controls | 64 |
| Customising search forms | 67 |
| File descriptions | 67 |
| Inserting query-dependent information | 68 |
| Language selection | 71 |
| Category filters | 71 |
| Date filters | 72 |
| Relevance threshold cutoff | 73 |
| Top terms | 74 |
| Widen search | 75 |
| Boolean queries | 75 |
| Searching from a static HTML page | 76 |
| Advanced information | 77 |
| Section 4 | |
| Reference | 79 |
| Glossary | 81 |
| Product information | 85 |
| Index | 87 |



Introduction

Muscat FX allows you to create indexes (databases) of documents and make them searchable via a simple configurable interface.

There are three components to Muscat FX:

- Index Management Program (imp) – This allows you to create indexes.
- Search front end (fx) – This allows end users to search the indexes you create.
- Index builder (indexer or msindexer) – This actually builds the indexes.

There are two versions of the Muscat FX software; one (Site Indexer) creates indexes of your *local* web server, the other (Multi-Site Indexer) can create an index of *remote* web sites using the WWW HTTP protocol.

Muscat FX uses a unique combination of techniques to search one or more indexes and home in on target documents:

- probabilistic – ranks documents according to the dynamic relevance of terms in the database
- word frequency – uses statistical techniques to rank documents according to how often query terms occur
- proximity – favours documents with search terms close to each other.



About this guide

This guide tells you how to install and set up the Muscat FX Site Indexer and Multi-Site Indexer products. It explains how to use the basic FX indexing and search facilities, and gives more advanced information on customising the FX products to suit your needs.

Muscat FX is currently available for the Unix and Windows NT platforms. This guide provides instructions for use on both platforms.

Who should read this guide?

This guide is aimed primarily at the person responsible for the day-to-day running of a web server.

Ideally, you should:

- be familiar with the Unix or Windows NT platform, as appropriate
- know where system files are stored, especially those relevant to web use
- have sufficient system privileges to be able to change access permissions and write to system files.

In addition, if you want to customise Muscat FX you should have some familiarity with Perl scripts and the HTML language.



How to use this guide

The guide is divided into four sections.

Section 1 – Getting started

This section explains how to get Muscat FX up and running.

- *Installing the software* – Tells you how and where to install the software, as well as listing the information you'll need to supply before you can complete the installation.
- *Basic configuration* – Tells you how to set up file mappings, host multiple sites and restrict access to index creation.

Section 2 – Basic use

This section explains how to create and search indexes.

- *Creating an index* – Tells you how to create a basic site or multi-site index, specify its appearance and behaviour, and define which types of file will be indexed.
- *Searching an index* – Tells you how to construct simple and advanced queries to search an existing index.

Section 3 – Advanced use

This section explains how to customise Muscat FX.

- *Running multiple indexes* – Tells you how to run a query over more than one database.
- *Customising FX* – Tells you how to enhance existing search form templates and customise the FX search engine.

Section 4 – Reference

This section contains the following chapters:

- *Glossary* – Gives definitions of terms used in this guide.
- *Product information* – Lists related Muscat products.
- *Index*.



About Muscat FX

The purpose of building an index is to allow someone browsing a site to search for all documents relevant to a given topic.

Most common search engines allow you to narrow down a search by entering multiple search terms. FX makes use of 'intelligent' search features, allowing you to target your query much more effectively:

- **Query expansion** – Muscat FX employs two mechanisms for personalising a query – the Expand and Improve functions. These add extra relevant terms to the search string which, unlike thesaurus-based, schemes, are dynamically tailored to a user's interests.
- **Word stemming** – Muscat FX stems words to increase the number of hits during a query. For example, when searching using any of the words 'index', 'indexing', 'indexes' or 'indexed', FX retrieves all documents starting with the string 'index'.
- **Proper names** – Muscat indexes proper names exactly as entered in the search string, if you enter them with initial capitals. For example, search for 'Stephen Hawking' rather than 'stephen hawking' to avoid retrieving all documents containing the word 'hawks'.
- **Changing the priority of weightings** – You can give a higher weighting to more important search terms by placing a plus (+) sign in front of them. Similarly, less important terms are preceded by a minus (-) sign. For example, to search for documents about computers but not about networks you would specify the following search string:
`+computer -networks`
- **Multiple document formats** – Muscat FX can index Word, PowerPoint, Excel and other common document formats using the Muscat Document Filter product. By default, Muscat FX indexes HTML, TXT and other ASCII formats. Muscat Limited also offers page-by-page indexing of Adobe Acrobat PDF documents, with word highlighting, as an additional Indexing Filter extension. See the chapter *Further product information*.
- **Index management program (imp)** – Muscat FX comes supplied with `imp`, a simple HTML-based program for managing indexes. `imp` also includes a set of example index templates illustrating different search facilities and graphic styles.



Section 1 ***Getting started***





Installing the software

This chapter tells you how to install and set up the FX software on the following platforms:

- Unix
- Windows® NT.



FX system architecture

This section gives a brief overview of the file and directory structure created by a typical Muscat FX installation.

CGI directory

The following programs are installed in the CGI directory:

| | |
|--------------------|--|
| <code>fx</code> | Muscat program allowing web users to search Muscat indexes of your website |
| <code>imp</code> | Administration program for setting up and customising indexes |
| <code>html</code> | Script for highlighting query terms found in HTML documents |
| <code>pdfhl</code> | Script for highlighting query terms found in PDF documents |

bin subdirectory

The following programs are installed in the `bin` subdirectory of the installation:

| | |
|---------------------------|---|
| <code>indexer</code> | Index website pages via the file system |
| <code>msindexer</code> | Index website pages via http |
| <code>update-index</code> | Updates indexes without having to carry out a complete rebuild |
| <code>remake-index</code> | Re-creates indexes from scratch |
| <code>mcl</code> | Muscat command line program to provide Muscat services to the indexer |

`fx` expects indexes to be in the `data` subdirectory of the installation directory. When setting up an index with `imp`, you will be given the option of creating the index in this directory. If you place them elsewhere, `fx` needs to be told where to find them:

- **Windows NT** – `imp` will add an entry to `muscatfx.ini` to point to the index location
- **Unix** – `imp` will create a symbolic link from the `data` directory to the actual location of the index.



Before starting the installation

Before you start to install the software you need to check you have up-to-date versions of system software and set the necessary write/access permissions.

System requirements

You'll need the following before installing the software:

- 2MB free disk space for the installation (you'll obviously also need disk space free for any databases you create)
- Perl version 5.002 or later (**Unix**), version 5.004_04 or later (**Windows NT**)
- (**Unix only**) An extraction/decompression utility, such as gnu tar.

Information you'll need

Before proceeding with the installation you will need to know the following:

- The location of 'webroot' – the root directory for HTML documents on your machine (also known as the 'primary document directory' in Netscape Suitespot).
- The location of `cgi-bin` – the directory where cgi scripts are installed
- The hostnames of any machines to be given access to the FX administration program (used to create, delete and control indexes)
- (**Unix only**) The name of the user under which httpd runs on your machine.

Make a note of this information, as you'll need it during the installation.

Muscat can be installed anywhere – the file `/etc/muscat-fx.cf` (**Unix**) or `winnt\muscatfx.ini` (**Windows NT**) points to the installation. Typically, Muscat is installed in `/usr/muscat` (**Unix**) or `c:\muscat` (**Windows NT**) but you can choose another location if you want.



Setting access permissions

You need to set certain access permissions before you start to install the software.

In particular, you need to ensure that the install script can:

- write to the directories where you want to place the software
- write to the `html` and `cgi` directories
- change ownership and access permissions on some files (so it may need to be run as superuser or administrator, otherwise you will have to perform manually any operations that fail).



Installing under Unix

This section explains how to install the Muscat FX software on a **Unix** platform.

Installation comprises the following steps:

- Extracting and uncompressing the software
- Running the install script.

Extracting and uncompressing the software

We recommend that you use `gnu tar` to extract and uncompress Muscat FX.

Type the following command:

```
tar -xvzf filename
```

The filename will depend on the flavour of **Unix** you have and whether you're installing the site or multi-site version of FX.

Running the install script

Once you've extracted the software you can begin the installation.

1. Start the install script by typing:

```
cd Muscatfx  
./muscat-install
```



For the most part, all you need do is follow the on-screen instructions and supply information when prompted. You can often simply accept the default values (supplied in square brackets) by pressing **Return**.

Note: You can quit the installation at any time by pressing CTRL-C.

The remainder of this section highlights the main steps in the installation process.

2. Enter the pathname for the installation, or press **Return** to accept the default
[/usr/muscat]
The script will check the existence of perl and associated modules. It will inform you if you need a more up-to-date version.
3. Once the checks are complete, press **Return** to accept the version of Perl found. Alternatively, specify a different pathname if you know there is a version of Perl somewhere else on your system.
4. Enter the pathname of the webroot directory, or press **Return** to accept the default
[/home/httpd/html]
5. Enter the pathname of the cgi-bin directory (where scripts are installed) or press **Return** to accept the default [/home/httpd/cgi-bin]
6. Enter the pathname of the cgi-bin directory when accessed from a web browser, or press **Return** to accept the default [/cgi-bin]
7. Specify the directory where Muscat graphics should be installed, or press **Return** to accept the default [/muscat]
This directory *must* be below the webroot directory.
8. Specify the hostnames of any machines to be given access to the Muscat administration program (`imp`)
Enter one or more IP addresses or full hostnames (e.g. marvin.stuff.com) with a new line after each one.
You can use '*' as a wildcard. For example, *.stuff.com will allow all machines in the stuff.com domain to run the administration program.



9. Enter a blank line (i.e. just press **Return**) to finish the list of hostnames.

You'll see the following messages:

```
Installing Muscat...
Installing fx...
Installing imp...
Installing graphics...
Modifying Muscat scripts...
```

10. Enter the name of the user that httpd runs as, or press **Return** to accept the default [http].

(The httpd process needs to have write permission in the `/usr/muscat/data` directory. This is achieved by having this directory owned by the httpd user.)

You'll see the following messages:

```
Changing ownership of the data directory to name ...
Saving settings in /usr/muscat/config.sh ...
Updating existing indexes ...
Installation complete!
```

You've now completed the installation. See *What next?* at the end of this chapter.



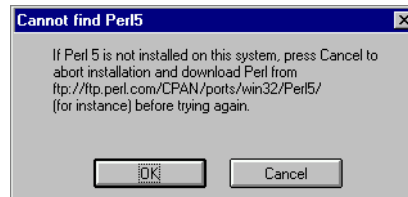
Installing under Windows NT

Here's how to install the Muscat FX software under Windows NT:

1. Double-click the install icon:

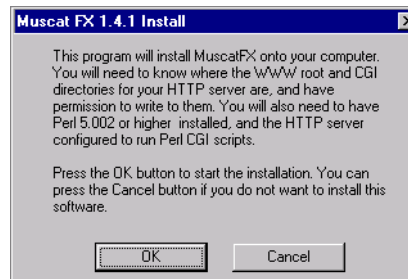


If you haven't got the correct version of Perl installed, you'll see the following message:



Click **Cancel**. Download and install the correct version of Perl before proceeding.

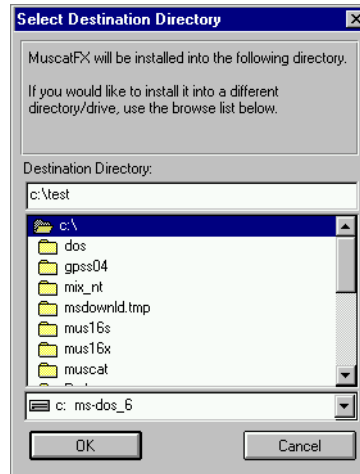
When you have got the correct version of Perl, you'll see the following screen:



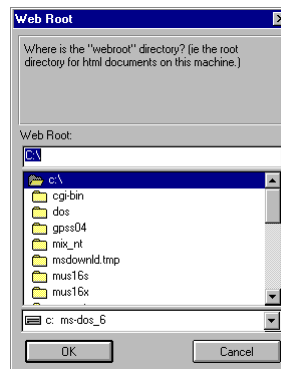
2. Click **OK**.



3. Select the destination directory where you want to install Muscat FX and click **OK**:

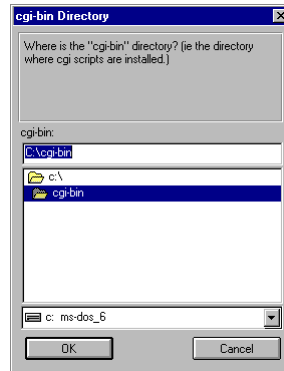


4. Select the location of the webroot directory and click **OK**:

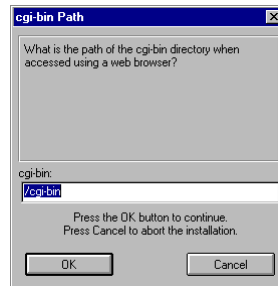




5. Select the location of the cgi-bin directory and click **OK**:



6. Select the location of the cgi-bin directory when accessed through a web browser and click **OK**:

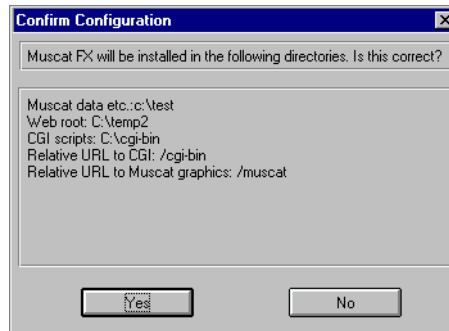




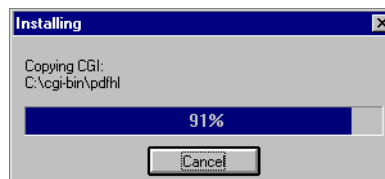
7. Select the directory where you want to install graphics and click **OK**:



8. Click **Yes** to confirm installation details:



You'll see a progress bar as the installation proceeds:





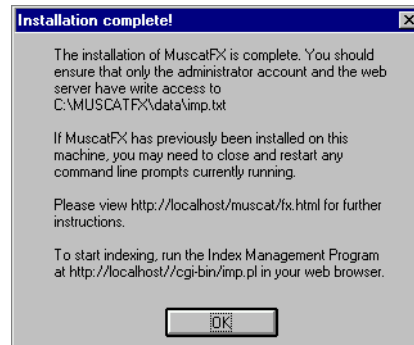
- Specify the hostnames of any machines to be given access to the Muscat administration program (`imp`).

Enter one or more full hostnames (e.g. `marvin.stuff.com`) with a newline after each one:



You can use `''` as a wildcard. For example, `*.stuff.com` will allow all machines in the `stuff.com` domain to run the administration program.

- Click **OK**.
- Click **OK** at the final screen to finish the installation:





What next?

Now you've installed Muscat FX, point your browser to the following URL:

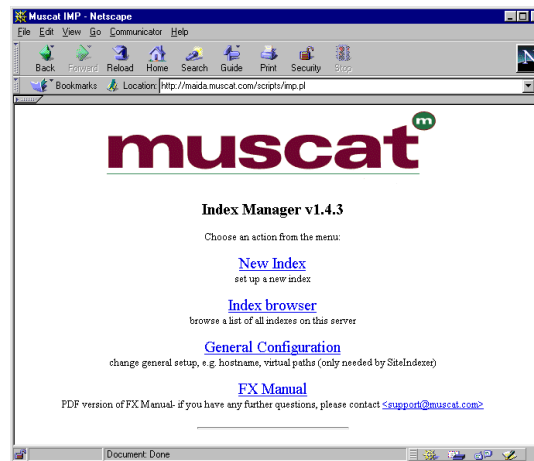
`http://my-host/my-cgi-bin/imp` (**Unix**)

or

`http://my-host/my-cgi-bin/imp.pl` (**Windows NT**),

where `my-host` is the hostname of your machine and `my-cgi-bin` is the location of your cgi-bin directory.

You'll see the main menu page for the Index Management Program (`imp`):



From here, you can:

- click on the **General Configuration** link to perform further configuration (this is described in the following chapter)
- click on the **New Index** link to start indexing (this is described in Section 2).



Basic configuration

This chapter explains how to configure your Muscat FX software for everyday use. You may already have specified some of the configuration options during the installation process, but this chapter tells you where configuration files are stored and how to edit them if necessary.

In particular, you can:

- set up file, URL and directory mappings
- set up your machine as a virtual host for multiple websites
- restrict access to the `imp` script (i.e. specify a list of machines which are allowed to create indexes).

The Muscat `data` directory (in the installation directory) contains two configuration files: `config.txt` and `imp.txt`. You can perform some of the configuration directly through your browser window, using the `imp` program, or you can edit the configuration files by hand.



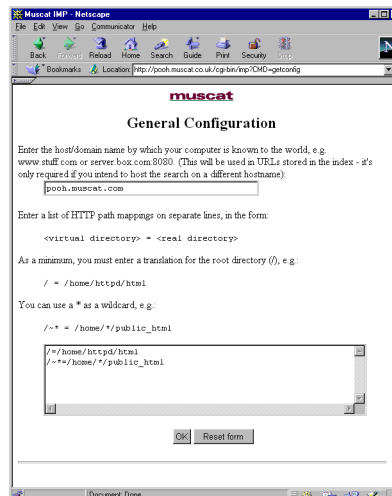
Setting up mappings

Note: This section only applies to the Site Indexer.

The `config.txt` file in the `data` directory controls file, URL and directory mappings. You can edit the file using any standard text editor, but it makes sense to make any changes via `imp` in a browser window.

The basic method for making configuration changes is as follows:

1. Click on the General Configuration link on the `imp` main menu to display the following page:



2. Enter mappings in the two boxes provided (see descriptions in the following two subsections).
3. Click **OK** to apply your changes and return to the main menu.



Running searches from another machine

By default, index searches are run on the machine on which they were created. However, you can query indexes of several machines from a single server. In order for this to work, you need to include the hostname of the machine on which the web pages are held in any URLs in the index.

To do this:

1. Display the General Configuration page as described above.
2. Enter the name of your machine in the first box:

Enter the host/domain name by which your computer is known to the world, e.g. `www.stuff.com` or `server.box.com:8080`. (This will be used in URLs stored in the index - it's only required if you intend to host the search on a different hostname):

3. Click **OK**.

Setting up directory mappings

The `config.txt` file also specifies the mappings between actual directories on your server and what a browser sees. For example, `/images` on a browser might actually point to `/ftp/pub/images` on your machine.

Mappings in the `config.txt` file must be in the following form:

```
virtual_path = real_path
```


Using the above example, you would need:

```
/images = /ftp/pub/images
```



To specify directory mappings:

1. Display the General Configuration page.
2. Enter all your directory mappings in the second box:



```
/* /home/httpd/html
/* ~* /home/*/public_html
```

Note: At the very least, you must specify a mapping for the root directory. By default, this box contains the mapping you specified during the installation. For example: / = /home/httpd/html

3. Click **OK**.

Using wildcards in directory mappings

You can use the * character as a wildcard in your directory mappings. For example, if your HTTP server is set up to map URLs like `/~phineas/` to `/export/home/phineas/public_html` you could have the following entry:

```
/* ~* = /export/home/*/public_html
```

Virtual hosts

If you are using a single machine to host two or more websites (*virtual hosts*), and want to create a combined index of the different sites, you can specify a full URL for the directory mapping, instead of the virtual path above.

For example, if you host `www.itchy.com` on `/www/itchy_home` and `www.scratchy.co.uk` on `/www/scratchy_pages`, you would edit the `config.txt` file (either by hand, or by adding the following lines to the mappings box on the General Configuration page):

```
http://www.itchy.com/ = /www/itchy_home
http://www.scratchy.co.uk/ = /www/scratchy_pages
```

This will cause full URLs to be stored in the index rather than partial ones.



Restricting `imp` access

The `imp.txt` file controls access to the `imp` script. It contains a list of hostnames or IP addresses that are allowed access to `imp`. For example, if you want the machines `wombat.marsupial.com` and `kangaroo.marsupial.com` to be able to use `imp`, the file should look like this:

```
wombat.marsupial.com  
kangaroo.marsupial.com
```

Alternatively you could put `*.marsupial.com` to allow any computers in this domain access to it.

For extra security, you can configure your HTTP server to make access to `imp` password protected. Consult your HTTP server documentation for details on how to do this.



Section 2 ***Basic use***



Creating an index

This chapter explains how to:

- create a site index for a local web server
(available with Site Indexer and Multi-Site Indexer products)
- create indexes for multiple sites
(only available if you have purchased the Multi-site Index product)
- build an index
- browse an index



Setting up the look and feel

Before you actually create an index, you need to specify where it will be created, specify a name for the index and define its appearance.

1. Point your browser to the following URL:

`http://my-host/my-cgi-bin/imp` (**Unix**)

OR

`http://my-host/my-cgi-bin/imp.pl` (**Windows NT**),

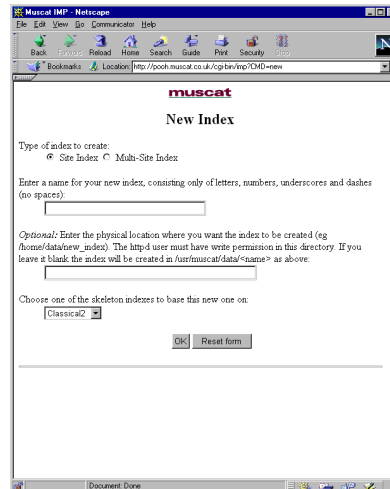
where `my-host` is the hostname of your machine and `my-cgi-bin` is the location of your cgi-bin directory. You'll see the main menu page for the Index Management Program (imp):





2. Click New Index.

You'll see the New Index page:



3. If you're using Multi-Site Indexer, click Site Index if you want to index your local server, or Multi-Site Index to index remote servers.
(If you're using *Site Index*, this option won't be available).
4. Enter a name for the index.
This can contain alphanumeric characters, full stops (.) and underscores (_) but not spaces.
5. Enter a physical location where the index will be created.
By default (if you leave the field blank), indexes are stored in the installation data directory. You might want to specify a different location if, for example, you installed

Muscat FX on a partition with limited space. If you do this, imp creates either a symbolic link to the new directory (**Unix**) or creates an entry in the muscatfx.ini file (**Windows NT**).

Note: Remember that the web server must have write permission in the specified directory.

6. Click the menu arrow and choose a *skeleton* for your index:

Classical2 ▾

The skeleton defines the look and feel for your index (e.g. colours, buttons and features). Muscat provides the following seven skeletons (you can modify them and create your own skeletons – see Section 3):

| Name | Graphic style | Features included |
|-------------------|------------------------------|-----------------------------|
| <i>Basic</i> | minimal | improve |
| <i>Classical</i> | plain | improve |
| <i>Classical2</i> | plain | improve & expand |
| <i>Granite</i> | square granite-style buttons | improve |
| <i>Marble</i> | green textures | expand |
| <i>Modern</i> | bold colours | improve |
| <i>Topterns</i> | Muscat EuroFerret search | improve & expand & topterns |

Here are a couple of example buttons:



Granite skeleton



Classical skeleton



7. Click **OK** to display the next page. The next page displayed depends on what type of index you're creating – see *Configuring a single-site index* or *Configuring a multi-site index* in the following section.

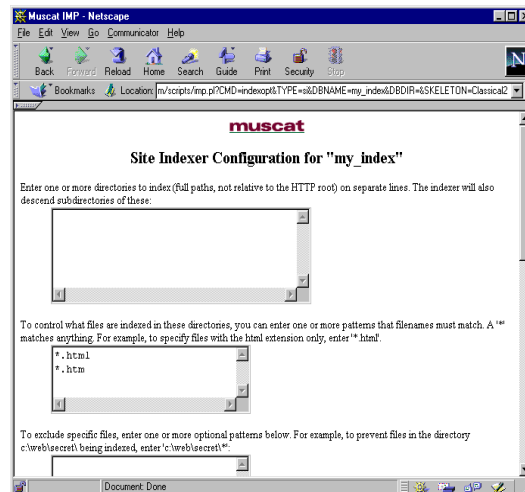
Configuring an index

This section explains how to specify what to include in your index. For instance, you can include or exclude specific documents, or follow symbolic links within indexed documents.

Configuring a single-site index

If you're using the Site Indexer, or if you specified the Site Index option in the previous section, you're now ready to create an index for your local server.

For example, if you've set up a single site index with the name `my_index`, the following page would be displayed:



1. In the first box, specify the directories you want to index. Muscat FX will index files in these directories, and in any subdirectories.



You must specify full pathnames (i.e. you can't enter names relative to the webroot directory), with each entry on a separate line.

2. In the second box, define the documents you want to include in the index. For example, `*.*` will include all documents. Don't leave this box blank – the index will be empty!

You can enter one or more patterns in this box, with each on a separate line. These match the entire pathname using standard shell syntax.

For example:

| | |
|-----------------------|---|
| <code>*.html</code> | matches all files with the <code>.html</code> suffix (but note that by default a new index will include <code>.htm</code> and <code>.html</code> files) |
| <code>/stuff/*</code> | matches all files in the <code>/stuff/</code> directory |

Note: If you have purchased a PDF or Office document filter, enter the correct extension (e.g. `.pdf`) to include them in the index.*

3. In the third box, use the same rules as above to specify documents you want to exclude from the index.

When you run the indexer, it will first reject any documents *not* in the Include list, then reject documents in the Exclude list.

4. (**Unix only**) Check the follow symbolic links option if you want to index virtual directories which are not directly below the Web root directory.

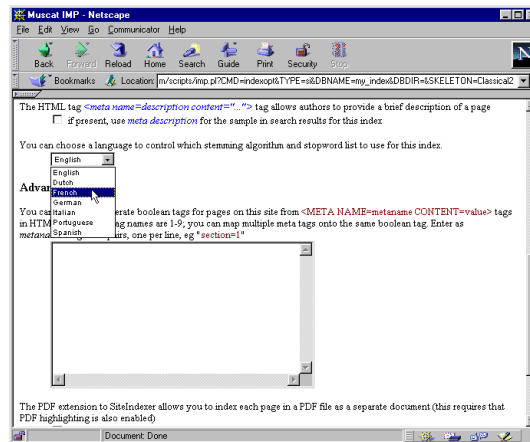
This assumes you have defined your mappings correctly (see *Basic configuration* in Section 1).

5. Check the use meta description... option if you want the indexer to display descriptive information included in some HTML documents.

Some authors use the following HTML tag to provide a short description of their documents:

```
<meta name=description content="...">
```

The indexer displays this content text in the results of a search, beneath the title of the document.

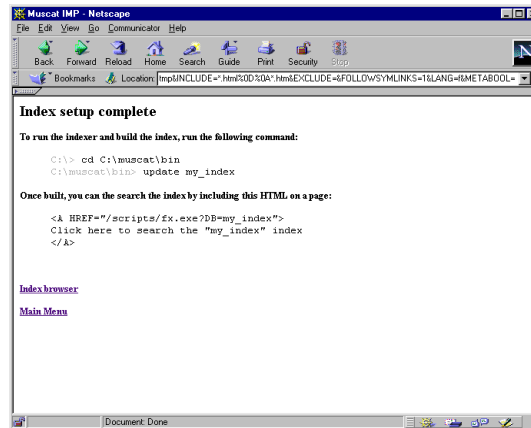


6. Select a language to control the stemming and stop word list, The supported languages include Dutch, French, German, Italian, Portuguese, Spanish with the default set at English.
7. Specify any special search categories in the Advanced options box.

While you're getting started, you can leave this box blank. You'll find details in *Advanced configuration options* at the end of this section.



8. Click **OK** to display the Index setup complete page:



You've now configured your index. See *Building an index* and *Using the index browser* for instructions on what to do next.



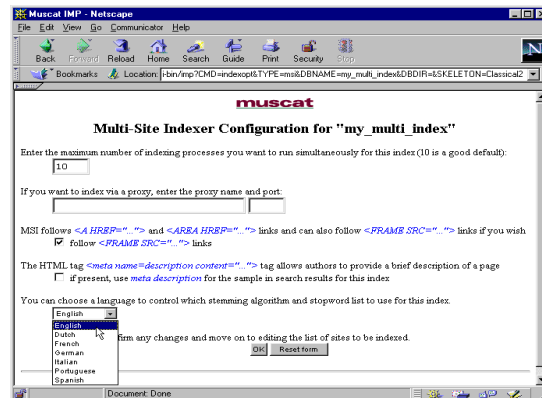
Configuring a multi-site index

This section explains how to configure indexes using the Multi-Site Indexer. You need to:

- set up some general parameters for the Multi-Site Indexer
- set up parameters for each site to be indexed.

General configuration

If, for example, you set up a multi-site index with the name `my_multi_index`, the following page would be displayed:



1. (**Unix only**) In the first box, enter the number of indexing processes you want to be able to run simultaneously.
 - Running more than one process allows the indexer to make better use of available bandwidth by retrieving several pages at once. If you have sufficient



bandwidth and processing power, allocate one process for each site you want to index.

- You may want to *reduce* the load the indexer places on the server by restricting the number of processes.
2. If you use a proxy server to access the sites you want to index, enter a server name and port number in the second box.
 3. If you're attempting to index frames-based sites, make sure the follow `<FRAME SRC="...">` option is checked (it is checked by default).

(If this option is unchecked, the indexer only follows links of the type `` and `<AREA HREF="...">`)

4. Check the *use meta description...* option if you want the indexer to display descriptive information included in some HTML documents.

Some authors use the following HTML tag to provide a short description of their documents:

```
<meta name=description content="...">
```

The indexer displays this content text in the results of a search, beneath the title of the document.

5. Select a language to control the stemming and stop word list. The supported languages include Dutch, French, German, Italian, Portuguese, Spanish with the default set at English.



6. Click **OK** to display the site configuration page:

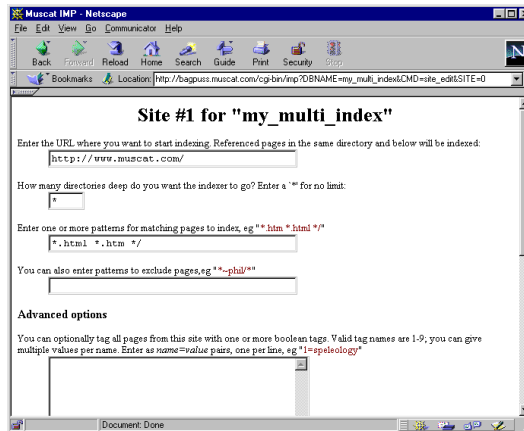




Site-specific configuration

You'll need to repeat the steps in this section for each site you're including in your index.

1. From the site configuration page, click *Add a New Site* to start configuring the index for your first site:



2. Enter the URL where you want indexing to start.

The indexer will follow links on this website at or below the same directory level.

For example, if you enter

```
http://www.prosimian.com/lemurs/index.html
```

the indexer will also pick up

```
http://www.prosimian.com/lemurs/language/
```



but not

```
http://www.prosimian.com/sifaka/
```

3. Specify the depth of directories to which you want to index, relative to the start URL.

For example, `*` specifies no limit, `0` limits the indexer to the same level and `1` allows indexing to one level below the start URL.

4. In the second box, define the documents you want to include in the index. For example, `*.*` will include all documents. Don't leave this box blank – the index will be empty!

You can enter one or more patterns in this box, with each on a separate line. These match the entire pathname using standard shell syntax.

For example:

```
*.html    matches all files with the .html suffix (note that by default a  
          new index will include .htm and .html files.
```

```
/stuff/*  matches all files in the /stuff/ directory
```

Note: If you have purchased a PDF or Office document filter, enter the correct extension (e.g. `.pdf`) to include them in the index.*

5. Use the same rules as above to specify documents you want to *exclude* from the index.

When you run the indexer, it will first reject any documents not in the Include list, then reject documents in the Exclude list.

6. Specify any special search categories in the *Advanced options* boxes.

While you're getting started, you can leave these boxes blank. You'll find details in the following section (*Advanced configuration options*).

7. Click **OK** to add this site to the list you want to index.



If you have filled in any of the configuration options incorrectly, you'll see a message telling you to correct them. Click **OK** to redisplay the previous page and amend your entries as appropriate.

You'll then see the *Edit sites* page:



This page has a table listing all the sites you have set up for indexing. You now have various choices:

- Click *Add a New Site* and repeat the above instructions for the next site you want to index
- Click on the URL for the site you've just set up if you want to amend any of the configuration parameters
- Click *Done* to view the *Setup Complete* page – this allows you to build the indexes – see *Building an index*
- Click *Index Browser* to review the current status of all your indexes – see *Using the index browser*.



Advanced configuration options

Documents in your site may use meta tags to categorise pages (for example, by department or product). You can use these categories to restrict a search.

This section explains how to specify categories in the Advanced options boxes on the index configuration page. Read *Customising search forms* in Section 3 to understand how to make use of this information when searching an index.

Advanced options (Site Indexer)

The text entry box in the Advanced options section allows you to make use of any additional information held in meta tags within the pages to be indexed. For instance, if you have tagged your pages as belonging to certain sections of the site, you will be able to restrict the search to return only documents from specific sections. Muscat FX allows you to specify up to nine different categorisations.

The meta tags must be of the following form:

```
<meta name=category_name content=value>
```

For example, if you had your pages categorised by department and product, a document might contain the following tags:

```
<meta name=department content=sales>  
<meta name=product content=thingummies>  
<meta name=product content=whatsits>
```

To set department as category 1 and product as category 2 you would enter the following in the Advanced options box:

```
department=1  
product=2
```

Below the text entry box is an option relating to the PDF extension. When selected the option enables indexing of each page as a separate document, this allows you to jump directly to the page that includes the matching words.



Advanced options (Multi-Site Indexer)

There are two boxes in the Advanced options for a multi-site index:

- The first box allows you to mark all the pages being indexed as having a specific value in one of the nine definable categorisation fields. This would allow you to restrict a search to just these pages as opposed to the entire index.
- The second box has exactly the same functionality as described in *Advanced options (Site index)*.

For example, suppose you want to index a site containing pages divided into Sales information and Marketing information. All the Sales pages are held under the directory

`http://www.oursite.com/sales/`

and all the Marketing pages are held in

`http://www.oursite.com/marketing/`

You can set up the index so users can just search the Sales information, just search the Marketing information, or search both. To do this, set up the index as two sites:

Set the URL for the first to be

`http://www.oursite.com/sales/`

and put the following string in the first Advanced Options box:

`1=Sales`

Set the URL for the second site to

`http://www.oursite.com/marketing/`

and put the following string in the first Advanced Options box:

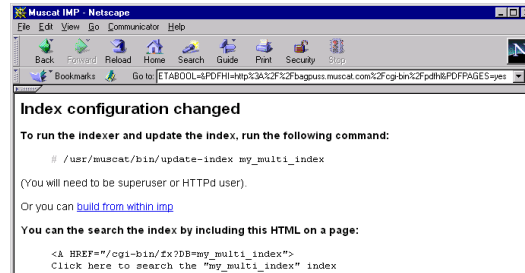
`1=Marketing`

When the Indexer is run it will attach the correct category code to each page and you will be able to customise the query form to allow users to search one or both of the sections (see *Customising search forms* in Section 3).



Building an index

Once you've set up and configured an index, you need to build it (i.e. run the indexer). *imp* automatically generates the command to run the indexer, and displays it in your browser window after you click **OK** in the *Site Indexer Configuration* page (Site Indexer), or after you click **Done** in the *Edit sites* page (Multi-Site Indexer):



The command has the form

```
update-index index-name
```

There are two ways to start the indexer:

- Copy the command into a terminal window and run it.

The indexer will display a message as each page is indexed, and a summary at the end.

- Click on the *build from within imp* link (**Unix only**).

In this case you won't see the status messages and summary as the index builds.



When the index has built, you can search it by pointing to the following URL (**Unix**):

```
http://my-host-name/my-cgi-bin/fx?DB=index-name
```

or (**WindowsNT**):

```
http://my-host-name/my-cgi-bin/fx.exe?DB=index-name
```

Updating and rebuilding indexes

You can update an existing index (if, for example, pages have been added, altered or deleted) by running the above update-index command again. In this case, the indexer

- removes records for files that have been deleted
- adds new files
- updates any files that have been modified.

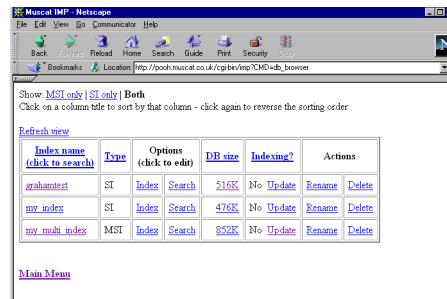
You can also rebuild an index from scratch using the following command:

```
remake-index index-name
```

Using the index browser

You use the Index browser to control and reconfigure indexes that you have previously set up.

Click on the *Index browser* link on the *Edit sites* page to display the browser:



- Information in the browser is displayed in a table, with a row for each available index.
- (Multi-Site Indexer only) At the top of the window you can click appropriate links to show multi-site indexes, site indexes, or both in the table.
- You can click the *Refresh view* link to update the browser information if indexing is currently in progress.
- Clicking on a column heading sorts the table by that field. Clicking again reverses the sort order.

For example, clicking on the *DB Size* column sorts indexes from smallest to largest size; clicking again sorts from largest to smallest.



Browser information

The columns in the Browser window contain the following information:

- **Index name** – The name of each index
- **Type** – (Multi-Site Indexer only)
SI = Site Index
MSI = Multi-Site Index
- **Options** – Contains two links (Index and Search) to allow you to access Indexing Controls and Search Controls – described later in this section
- **DB Size** – If there is a database file present, this displays the size of the file; click on this to see information on the database contents
- **Indexing? (Unix only)**– Status of indexer
Yes = index is currently being built
No = index is complete
Click *Update* to rebuild the index
- **Actions** – Contains two links which allow you to Delete or Rename the index.

Note: If the index is currently being built, you can't delete or rename the index – these options are not displayed.



Database information

Clicking on a link in the DB Size column displays more information about that particular index:

The screenshot shows a web browser window titled 'Muscat IMP - Helixpage'. The address bar contains the URL 'http://p00h.muscat.co.uk/cgi-bin/vmp/CMO-indent.html?DBNAME=gahantest'. The main content area displays a table with the following data:

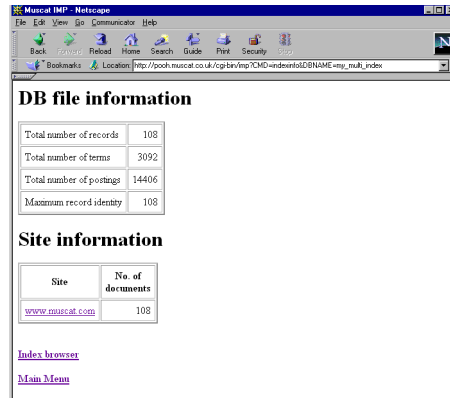
| DB file information | |
|--------------------------|------|
| Total number of records | 80 |
| Total number of terms | 3156 |
| Total number of postings | 5496 |
| Maximum record identity | 80 |

Below the table, there are two links: [Index browser](#) and [Main Menu](#).

- **Total number of records** – There is one record in the database for each page indexed.
- **Total number of terms** – Each term represents a unique, searchable word found in the source pages. There are also terms for other pieces of information, such as the URL for each page.
- **Total number of postings** – A posting is an instance of a term in the index (e.g. if each term occurred three times within the index, there would be three times as many postings as terms).
- **Maximum record identity** – Each record in a Muscat index is allocated an identity number when it is added. The maximum value will be the same as the number of records until records are deleted from the index. The difference between the number of records and the maximum record identity equals the number of records that have been deleted.



- **Site Information** (Multi-Site Indexer only) – An extra table appears at the bottom of the page showing the number of documents at the currently-displayed site:



Indexing controls

Once you've created an index, you can edit some of the indexing parameters. To do this:

1. Click the *Index* link in the *Controls* column of the *Index Browser* page.

You'll see the *Index Control* page, which is similar to the *New Index* page. You can edit the following indexing parameters:

- The list of directories to be indexed
- The Include and Exclude lists.

Please see *Configuring an index* earlier in this chapter for details.



2. Once you're happy with the parameters, click **OK**.
3. Run the `update-index` or `remake-index` command as appropriate.

Note: If you've made lots of changes to the index, it's usually best to remake it.

Search controls

You can use `imp` to control some of the behaviour of the FX search program, on a per-index basis. For example, you can alter the maximum number of results displayed on the query page, or the number of words displayed in the Improve and Expand lists.

You can access these controls by clicking the *Search* link in the *Controls* column of the *Index Browser* page.

Please see the appropriate section in *Customising FX* in Section 3 for details.



Searching an index

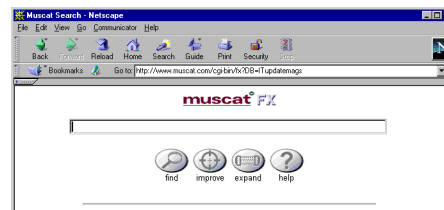
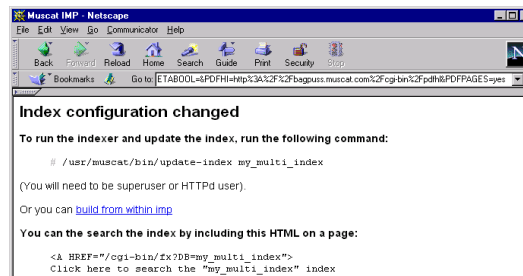
This chapter gives a brief description of the Muscat FX search engine. The front end has been designed to be fairly straightforward to use, and easy to customise (see later in this guide for details on customising search forms).

Displaying a search form

There are two ways you can give a user access to a search form for a particular index:

- Provide a standard URL to link to the relevant query page (this is the usual method)
- Include a static HTML form in a web page (please see Section 3 of this guide for details).

When you've built an index, `imp` provides the URL you need to include on your web page to allow users to access the search form:





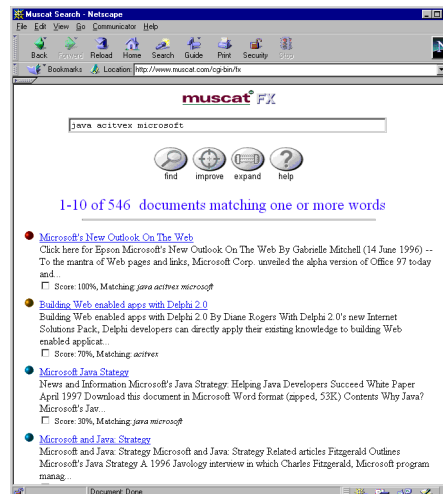
Submitting a query

Basic search

To submit a query:

1. Enter one or more search terms (separated by spaces) in the text box.
2. Click the **Find** button or press **Return**.

FX will search the index and display the first of any Results pages:





This page displays the following information:

- Number of documents found
- Score GIFs – Graphics down the left hand edge of the page indicating a document's relevance
- Document titles, which are clickable
- Document description – Either the first few words of the document, or (if applicable) the text from the `<meta name=description content="...">` tag.
- Relevance check-boxes (checking a box marks that document as being relevant to any further search operations such as Expand or Improve).
- 'Hit' navigation buttons

A Results page will only show a subset (10, by default) of documents matched in the search. To view another page-worth of hits, click the *Previous* or *Next* buttons at the bottom of the page, or click on one of the numbers to the right.

3. Click on a document title to view it.

Expanding a query

To refine your search:

1. Place a check next to any documents you're interested in:

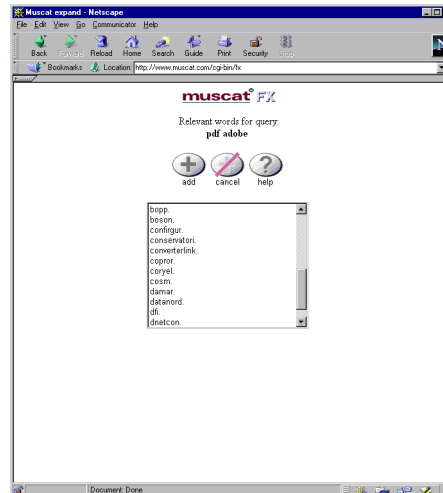
Score: 30%, Matching: *java microsoft*

2. Click the **Expand** button at the top of the *Results* page:





3. FX will display a list of terms that you could add to the query to improve the quality of the results:



4. Select a term (Shift-click to select multiple terms) and click the Add button.
(Click **Cancel** to clear any selections and return to the previous Results page.)
FX will search the index using your expanded query list and display the results.

Improving a query

The *Improve* function is an automatic version of *Expand*.

If you click the **Improve** button, FX automatically adds into the query the terms most likely to improve the search (i.e. you don't need to select relevant terms from a list).



Section 3 ***Advanced use***





Running multiple indexes

Muscat provides a mechanism for running a query over more than one database. Basically, the DB file in the `d/` subdirectory of the index directory is replaced by a list of DB files (called a 'dlist').

Note: Although it is not possible to search across a dlist which combines multi-lingual indexes, it is possible (without having to declare a dlist) to provide a search in different languages (see Customising search forms: Language selection for details).

To create a combined index (**Unix**):

1. Create a new index directory, either by using `imp`, or by hand. The appearance and behaviour of the combined index will be controlled by the files in this new directory.
2. Edit the file `run_db` in the new index directory to contain the following text:

```
c i
enquire dlist DBS pspec t/pspec with t/espec
```
3. Create a new file `DBS` in the new index directory and populate it with an entry for each database you want to search over.

Each entry should have the form:

```
db file path_to_index_directory/d/DB
```

Here is an example `DBS` file:

```
db file /usr/muscat/data/main_index/d/DB
db file /home/fred/Muscat/my_index/d/DB
db file /usr/muscat/data/archive/d/DB
```

Note: It is recommended that a maximum of 20 databases are joined together, if you require more than this contact support for advice.

4. You can now search the combined index with FX, using a URL of the form:
`http://my-host/my-cgi-bin/fx?DB=combined_index`



5. You can optionally give Muscat more space if you are dlisting a number of indexes. The file `/etc/muscat-fx.cf` controls how much memory Muscat has when running. The default value is 800 (Kbytes). The file will, by default, contain a line like this:

```
/usr/muscat 800
```

assuming, of course, that you've installed Muscat into the `/usr/muscat` directory.

To create a combined index (**Windows NT**):

1. Create a new index directory, either by using `imp`, or by hand. The appearance and behaviour of the combined index will be controlled by the files in this new directory.

2. Edit the file `run_db.txt` in the new index directory to contain the following text:

```
c i
enquire dlist DBS pspec t/pspec with t/espec
```

3. Create a new file `DBS.txt` in the new index directory and populate it with an entry for each database you want to search over.

Each entry should have the form:

```
db file path_to_index_directory\d\db
```

Here is an example `DBS.txt` file:

```
db file c:\muscat\data\main_index\d\db
db file t:\home\fred\muscat\my_index\d\db
db file c:\muscat\data\archive\d\db
```

Note: It is recommended that a maximum of 20 databases are joined together, if you require more than this contact support for advice.

4. You can now search the combined index with FX, using a URL of the form:

```
http://my-host/my-cgi-bin/fx.exe?DB=combined_index
```
5. You can optionally give Muscat more space if you are dlisting a number of indexes. The file `winnt\muscatfx.ini` controls how much memory Muscat has when running. The file will include text like this:

```
[startup]
directory=c:\muscat
space=1200
```

The directory is the path to where Muscat is installed. Space is in Kbytes and defaults to 1024 if not specified.



Customising FX

This chapter tells you how to:

- control the behaviour of the FX search program on a per-index basis
- customise the appearance and behaviour of search forms.



Search controls

You can use `imp` to control some of the behaviour of the FX search program, on a per-index basis.

- Click the **Search** link in the *Options* column of the *Index Browser* window.

The resulting page allows you to alter the following items:

Graphics directory

This allows you to choose which directory (relative to the web root directory) will be used for the score and page link graphics on the query page. This does not affect button graphics, which you need to adjust by hand. See *Customising search forms* later in this chapter.

Dimensions of score gifs

This enables you to enter the height and width of the score images to speed up loading, or to alter their appearance.

Dimensions of page gifs

This enables you to enter the height and width of the page link images, to speed up loading, or to alter their appearance.

Number of hits per page

This controls the maximum number of results displayed on the query page. The default is ten.

Number of words to add for Improve

This sets the number of words that will be added to the query when a user clicks the **Improve** button. The default is five.



Number of words to display in Expand list

This sets the number of words to display when a user clicks the **Expand** button. The default is thirty.

Type of proximity reordering to use

This allows you to select one of the three types of proximity reordering available within Muscat:

- Tight – Increases the weighting of a document if pairs of words in the query appear very close together within the document.
- Loose – Increases the weighting of a document if pairs of words in the query appear fairly close together within the document.
- Phrase – Increases the weighting of a document only if all of words in the query appear close together within the document.

Number of documents to reorder for term proximity

This controls how many documents (from the top of the results list) to re-rank according to term proximity (i.e. how close together the query terms are in the document).

Because there is a slight additional overhead in using proximity reordering, this number should not be too large (i.e. more than 200). If this field is blank, no proximity reordering will be used.

Hide Expand checkboxes on query screen

This prevents checkboxes next to captions from being displayed, in effect disabling the Expand or Improve functions (they depend on documents being marked). This option will simplify the query screen. Note that you'll still need to edit the query forms (see *Customising search forms* later in this chapter) to remove the Expand or Improve command buttons.



Enable logging of search requests

This creates log files detailing each query run on the system. The two log files will be called `fx.log` and `query.log`. They will be placed in the Index directory (e.g. `/usr/muscat/data/my_index_name`).

| | |
|------------------------|---|
| <code>fx.log</code> | Contains information about who accessed the system, when, and what actions they performed. It is held in Common Log Format used by most web servers and can be analysed using any of the tools available for creating statistics on web server usage. |
| <code>query.log</code> | Contains a list of queries run on the system, with one query displayed on each line of the log file. |

Highlighting of words in HTML documents (Site Indexer only)

Enabling HTML highlighting will cause the HTML pages to be retrieved by a Muscat script called `html` and processed to highlight the matching terms before displaying the page to the user. The words can be highlighted by making them bold, underlining them or changing their colour.

You need to take care with this option because it bypasses the HTTP server, possibly giving users access to restricted pages. You can reduce the risk by setting two variables at the start of the `html` script which limit the pages that it will accept to highlight. See the comments at the beginning of the `html` script for details.

Highlighting of words in PDF documents (Site Indexer only)

Enabling PDF highlighting will cause the PDF pages to be processed by a Muscat binary called `pdfhl`. `pdfhl` scans the PDF file for the locations of matching terms, this information is used by the PDF plug-in on the local browser to highlight the appropriate words.

Unlike HTML highlighting PDF highlighting does not bypass the HTTP server and therefore incurs no security risk.



Customising search forms

The appearance of FX's search forms is defined for each index by a set of files in the `index/html/` directory. The following subsections tell you:

- what each file contains
- how to use Muscat-specific `\NAME` tags to insert query-specific information into the files
- the recommended procedure for customising search forms.

File descriptions

The following files control a search form's appearance:

| | |
|---------------------------|--|
| <code>query</code> | Controls the appearance of the main page, with the query text box, Search, Improve and Expand buttons, and the caption list. |
| <code>expand</code> | Used by the Expand function (but not Improve). Displays a list of words for the user to choose from. |
| <code>expand_error</code> | Displays an error message if the user tries to Expand or Improve without any documents selected. |
| <code>muscat_error</code> | Displays an error message if an internal Muscat error occurs. |



Inserting query-dependent information

FX pages are designed to be very similar to standard HTML to make them simple to customise. However, text needs to be inserted into the pages when a query is run. To make this possible, FX recognises Muscat-specific tags within the HTML files which must be replaced. These tags have the format `\NAME`. The standard tags used within the example styles provided are:

| | |
|-----------------------|--|
| <code>\PROB</code> | Inserts the current query text. |
| <code>\STATS</code> | Displays the number of documents found by the query. |
| <code>\HITS</code> | Inserts the caption list. |
| <code>\HITLINE</code> | Will ignore this line when the caption list is empty. |
| <code>\PREV</code> | Creates a button to display the Previous page of hits, if there is one. The format is controlled by text following the tag (see below). |
| <code>\NEXT</code> | Creates a button to display the Next page of hits, if one exists. |
| <code>\PREVOFF</code> | Displays an image if this is the first page of hits. Useful to display a greyed out button. See below for details. |
| <code>\NEXTOFF</code> | Displays an image if there are no more hits to display. Useful to display a greyed out button. See below for details. |
| <code>\PAGES.x</code> | Creates a set of buttons to jump to a specific page of hits. <code>x</code> can be either <code>T</code> or <code>G</code> . <code>T</code> displays standard HTML buttons. <code>G</code> uses customisable graphics (see below). |
| <code>\SAVE</code> | Inserts essential hidden fields which encode the 'state' of the Muscat query. This must be present in the files. |
| <code>\TERMS</code> | For the Expand form, inserts the relevant words found by the Expand process. |
| <code>\TOPDOC</code> | Used in the Expand form. Inserts the number of the first caption displayed in the previous query page. |



`\SCRIPT_NAME` Inserts the name of the fx binary into the form. This allows the software to continue to work if you need to rename the fx binary. For example, use the following in the query page:

```
<FORM METHOD=POST ACTION="\SCRIPT_NAME">
```

In general, the best approach to creating a custom set of search forms is to start with one of the examples supplied and alter it incrementally, checking at each stage that it still works as expected.

The `\PREV`, `\PREVOFF`, `\NEXT`, `\NEXTOFF` and `\PAGES.x` tags are slightly complicated, and work the following way:

The format of the Previous or Next buttons is controlled by the text following the tag, which is read up to the end of the line, or until the next slash (`\`) character. For a simple HTML button it should look like this:

```
\PREV TYPE="submit" VALUE="Previous"
```

while a graphical button would be created by the following:

```
\NEXT TYPE="image" SRC="/gifs/next.gif" BORDER=0 WIDTH=30 HEIGHT=30
```

(Note that this should all be on a single line).

Displaying greyed-out buttons

If you want a greyed-out button to appear on the first or last page of hits you can use the `\PREVOFF` and `\NEXTOFF` tags.

For example:

```
\NEXTOFF TYPE="image" SRC="/gifs/next-grey.gif" BORDER=0 WIDTH=30 HEIGHT=30
```

Creating buttons to display specific 'hits' pages

The `\PAGES.T` tag creates a set of simple buttons to display specific pages of hits, while the `\PAGES.G` tag creates graphical buttons. The graphics for these are obtained from the graphics directory configured with the



Search Control page, and should be called `page-n.gif` where `n` is 1 to 10. The dimensions of these GIFs should be uniform; you can enter them into the *Search Control* page to speed up page display.

Customising 'hit relevance' graphics (score GIFs)

You can also customise the graphics used to indicate the relevance of each hit. These are held in the same directory as the page buttons above. They are called `score-n.gif`, where `n` is 0 to 10; you can enter the dimensions of the score GIFs on the *Search Control* page.

Running a Query, Expand or Improve

Buttons for running a Query, Improve or Expand are created with standard HTML form tags `SUBMIT` or `IMAGE`. The meaning interpreted by FX is based on the tag name, not its value. These are the valid names and their function:

| | |
|---------------------|---|
| <code>EXPAND</code> | Creates an Expand list based on documents which have been marked. |
| <code>AUTOEX</code> | Automatically Expands (or Improves) the query |
| <code>Fnn</code> | Shows the results from document number <code>nn</code> |
| (no name) | Show results from first document |

Here are some examples:

- Query Submit button (text)
`<INPUT TYPE="SUBMIT" VALUE="Click me to search">`
- Query Submit button (graphical)
`<INPUT TYPE="IMAGE" SRC="/search.gif" BORDER=0>`
- Improve button (graphical)
`<INPUT TYPE="IMAGE" NAME="AUTOEX" SRC="/improve.gif" BORDER=0>`
- Cancel button on Expand page (text)
`<INPUT TYPE="SUBMIT" NAME="F\TOPDOC" VALUE="Back">`

Notice how the last example uses the `\TOPDOC` tag to get the appropriate document number.



Language selection

When building an index it is possible to select which language stemming algorithm and stop list you want the index to use (see *Creating an index* in Section 2). An index can have only one language associated with it and it is not possible to combine multi-lingual indexes together with a dlist. But by creating an index in each language (from data of that same language) it is possible to give the user the option to use the search facility in a range of languages.

For example, three indexes are created, one called “french” (indexed from French data and using the French stemmer), one called “english” (indexed from English data and using the English stemmer) and one called “german” (indexed from German data and using the German stemmer). A HTML page is created to allow the user to select a language, the page includes this code:

```
<A HREF="/my-cgi-bin/fx?DB=french">Search in French</A><BR>
<A HREF="/my-cgi-bin/fx?DB=english">Search in English</A><BR>
<A HREF="/my-cgi-bin/fx?DB=german">Search in German</A><BR>
```

When “Search in French” is selected FX is run with the “french” index.

Category filters

If the pages within the Index were categorised when they were created, you can get Muscat to generate an HTML `SELECT` tag containing all the different possible values for that category. A user can then select one or more options from this list to restrict the pages that are searched.

There are nine fields available to store different categories within the Index. These are set up using the *Advanced Options* boxes as described in the section on advanced configuration options in *Creating an index* in Section 2.

For example, to insert a `SELECT` tag to filter on category 1, insert the following code into the query page:

```
<SELECT NAME=B SIZE=3 MULTIPLE>
<OPTION VALUE=" " > -----Any-----
\BOOL.F1
</SELECT>
```

The `\BOOL.F1` tag should be replaced by `\BOOL.F2` for category 2, and so on.



Date filters

Date filters can be used to restrict a search by document date stamp.

Selection by year, month and day

Muscat includes three date fields to filter documents by year, month and day. Use the `\BOOL.Y` tag to filter on year, the `\BOOL.M` tag to filter on month and the `\BOOL.D` tag to filter on day.

For example, to provide three `SELECT` lists (one for year, one for month, one for day) which enable the user to restrict by date insert the following code into the query page:

```
<SELECT NAME=B SIZE=3 MULTIPLE>
<OPTION VALUE=" "> -----Any-----
\BOOL.Y
</SELECT>
<SELECT NAME=B SIZE=3 MULTIPLE>
<OPTION VALUE=" "> -----Any-----
\BOOL.M
</SELECT>
<SELECT NAME=B SIZE=3 MULTIPLE>
<OPTION VALUE=" "> -----Any-----
\BOOL.D
</SELECT>
```

Date range by text entry

To provide the user with text entry boxes to specify a date range, use the `\DATE1` tag (for start date) and `\DATE2` tag (for end date).

The following code (inserted in the query page) provides two boxes where date can be entered in the form of YYYYMMDD:

```
Start date
<INPUT NAME="DATE1" VALUE="\DATE1" SIZE=8>
End date
<INPUT NAME="DATE2" VALUE="\DATE2" SIZE=8>
```



Within number of days

To restrict a search to a number of days within today's date use `DAYSMINUS` in combination with `DATE1`.

For example, to provide a list which restricts the search to those documents dated anytime or within the last day, week, month or year insert the following code in the query page:

```
<SELECT NAME=DAYSMINUS>
<OPTION VALUE="0" \DMSELECT000>dated anytime
<OPTION VALUE="2" \DMSELECT002>dated within the last day
<OPTION VALUE="7" \DMSELECT007>dated within the last week
<OPTION VALUE="31" \DMSELECT031>dated within the last month
<OPTION VALUE="365" \DMSELECT365>dated within the last year
</SELECT>
<INPUT TYPE=hidden NAME="DATE1" VALUE="TODAY">
```

The `\DMSELECT` tag ensures a selected option remains selected.

Relevance threshold cutoff

Normally Muscat returns all the hits that match any of the words in the query. This can result in the hit list containing documents at the end which are only 2–3% relevant. You can put an HTML `SELECT` tag on the query page to allow a user to select a relevance threshold. All documents marked as less relevant than this threshold will not be displayed.

For example, if the threshold is set to 100%, only the documents matching *all* the words will be displayed.

To display the threshold `SELECT` tag, insert the following code into the query page:

```
<SELECT NAME=THRESHOLD>
<OPTION VALUE="0" \TSELECT000>0%
<OPTION VALUE="25" \TSELECT025>25%
<OPTION VALUE="50" \TSELECT050>50%
<OPTION VALUE="75" \TSELECT075>75%
<OPTION VALUE="100" \TSELECT100>100%
</SELECT>
```

You can use any threshold value from 0-100 simply by changing the value and the number in the `\TSELECT` tag.



Top terms

Top terms puts a list of terms that FX deems relevant to the current query on the search page. Clicking on a check box next to any of these terms inserts it into the query box so it will be added to the query next time it is run.

Top terms uses Javascript to achieve these results. The following Javascript code needs to be at the top of the query form:

```
<script language=javascript>
<!--
function C(c) {var i, o;
o = document.P.P.value;
if (c.checked){
document.P.P.value = o+" "+c.value;
} else {
o = " "+o+" ";i = o.lastIndexOf(" "+c.value+" ");
if (i!=-1) {
document.P.P.value =
o.substring(1,i)+o.substring(i+c.value.length+1,o.length-1);
}}
// -->
</script>
```

You also need to give the form a *NAME* attribute. Change the *FORM* tag on the query page to read:

```
<FORM NAME=P METHOD=POST ACTION="\SCRIPT_NAME">
```

Finally, where you want the terms to appear, insert some HTML code similar to the following:

```
<TABLE BORDER=0><TR><TD BGCOLOR="#ccffcc">
<FONT FACE=helvetica,arial>
\TOPTERMS
</FONT>
</TD></TR></TABLE>
```

The *Top terms* skeleton distributed with FX uses the *TOPTERMS* tag – you can use this as a reference when creating your own pages.



Widen search

If you have a connection to the Internet then you can insert a button on your query forms that sends the current query to the *Muscat EuroFerret* index of European web sites. Clicking on the *Muscat Euroferret* link will take you to the *Euroferret* site.

To include such a link to your page you need to insert some HTML into the query page.

For example:

```
<A HREF="\WIDEN">  
<IMG ALT="Web Search" SRC="/muscat/oval/web.gif" HEIGHT=56 WIDTH=60  
BORDER=0>  
</A>
```

FX replaces the `WIDEN` tag with a link to the *Muscat Euroferret*.

Boolean queries

Within FX it is possible to perform traditional boolean queries without using probabilistic methods to rank the resulting documents. To turn on the boolean functionality there needs to be a form tag called `TREATASBOOL` on the query page.

The tag can be placed in the HTML form as a check box to allow users to switch between probabilistic and boolean queries. For example, you could put the following in the query form:

```
<INPUT TYPE=checkbox NAME=TREATASBOOL \BOOLQUERY>
```

The `BOOLQUERY` tag is used to maintain the state of the checkbox.

When using boolean queries users can use the following operators between words in the query:

```
AND  
OR  
NOT
```

If there is no operator between two words then `AND` is assumed. Round brackets can be inserted around parts of the query to control the order in which the operators are evaluated.



For example:

```
(Tony OR Cherie) AND Blair
```

will return documents containing Tony and Blair as well as documents containing Cherie and Blair.

```
Tony OR (Cherie AND Blair)
```

will return documents containing just Tony as well as documents containing Cherie and Blair.

```
(Tony Blair) OR (Cherie Blair)
```

will return documents containing Tony and Blair as well as documents containing Cherie and Blair.

Searching from a static HTML page

You can run a query and get results straight from a static HTML page without first having to get a blank Query page, by sending FX a field called P containing the search text.

Here is an example form, which you could include in your home page:

```
<FORM ACTION="/my-cgi-bin/fx" METHOD="POST">
```

Search for:

```
<INPUT NAME="P" SIZE=30>  
<INPUT TYPE="SUBMIT" VALUE="Find">  
<INPUT TYPE="HIDDEN" NAME="DB" VALUE="my-database-name">  
</FORM>
```



Advanced information

This section describes some of the more complicated areas of Muscat FX. These are not needed for everyday use, so feel free to skip this section.

pspec file

Each index has its own `pspec` file, in the `t/` subdirectory of the `index` directory. This file controls the appearance of captions returned by a search. It only affects the main text portion of the caption, not the score GIF or information line. By editing this file, you can change the appearance of the captions (to a limited extent). It's quite easy to make a mistake which will stop the index working, so back up your work frequently (you can always go back to a `pspec` from a supplied skeleton in an emergency).

The `pspec` file consists of a set of directives. Lines starting with a `d` format the printing of each field, and have the syntax:

```
d *fieldname (a) symbols;
```

where `symbols` is a list consisting of the following elements:

| | |
|--------------------|--|
| <code>+</code> | prints the value of the field |
| <code>'foo'</code> | prints the string <code>foo</code> |
| <code>g0</code> | prints a newline (not an HTML <code> </code> tag) |
| <code>k</code> | suppresses the output of the field |

These elements can be in any order, and repeated if required. For example:

```
d *url (a) '<A HREF=' + '>' g0;
```

if the `*url` field contains `http://www.opossum.com/pogo.html` then the output would be:

```
<A HREF=http://www.opossum.com/pogo.html>
(newline)
```

The order fields are printed and can be changed with the `o` directive:

```
o *rec *field1 *field2 *field3 ...
```



The fields you might want to display are:

| | |
|-----------------------|-------------------------|
| <code>*url</code> | The URL of the page |
| <code>*caption</code> | The title of the page |
| <code>*sample</code> | A sample of the content |
| <code>*host</code> | Your hostname |

This is not a complete description of the `pspec`, which is beyond the scope of this guide.

espec file

Each index has an `espec` file in the `t/` subdirectory of the index. This file contains configuration data for the search engine. Most of it should not be altered, however there are a couple of things you can change:

| | |
|----------------------|--|
| <code>wdf</code> | If this appears on a line by itself, it switches Within Document Frequency ranking on, which ranks documents with more occurrences of a word higher than documents with fewer occurrences of the word. You can switch it off by commenting the line out, with a backslash (<code>\</code>) at the beginning of the line. |
| <code>reverse</code> | This causes documents that were added most recently to appear highest in the caption list (within the overall ranking). Commenting it out will cause earlier documents to appear first. Note that this does not depend on the modification date of the HTML files themselves. |



Section 4 Reference





Glossary

expand

Expand functionality takes a set of documents marked by the user as relevant to their query and produces a list of words that could be added to the query to improve the quality of the results.

HTTP

Hyper-Text Transfer Protocol. This is the protocol used to transmit web pages over the Internet or Intranet to a browser.

imp

The Index Management Program. This program controls the setting up and management of Muscat indexes.

improve

Improve functionality works in a similar manner to expand except the system automatically adds into the query the words most likely to improve the quality of the results.



information line

This is the line of text displayed on the results page that details the number of documents that have been found that match the current query.

installation data directory

When you install muscat you are requested to enter a directory in which to install the program files, this directory has a sub-directory called data which is the default location for indexes. This is the installation data directory.

Muscat document filter

This is an add on product to the Muscat FX product that allows you to index many other third party document formats, including MS Word, MS Excel and Adobe Acrobat.

page gif

Page gifs are a set of gifs on the results page that provide direct links to the first 10 pages of results. If there are less than 10 pages of results then less links are displayed.

probabilistic

Muscat ranks documents based on the probability that they are relevant to the user.

proximity

The distance between query terms in a document can also affect the document's position in the results.

***relevance***

The relevance of a document represents how closely it reflects the concept about which the user is trying to retrieve information.

score gif

A score gif is displayed beside each entry in a list of results. There are 10 different score gifs used to indicate different levels of relevance.

skeleton

The name given to one of the base directory structures from which each Muscat index directory is generated. Each skeleton directory can have its own look and feel as well as differences in functionality.

stemming

Stemming is the process of reducing all cases and forms of a word to a single root form. This means that plurals of words and different cases of regular verbs are all treated like the root word.

superuser

The superuser on a system has privileged access to all areas of the computer's filing system.

virtual hosts

Virtual hosts are a method of hosting several web sites on one physical machine. You should consult the documentation for your web server for more details.



word frequency

The number of times a word occurs within a document is taken into account when ranking results.

write permission

Some file systems only allow privileged users access to some sections. Muscat requires the system to be set up so that the imp program can write to certain directories.



Product information

Product support

If you have a Support and Maintenance contract you can obtain technical support by contacting:

`support@muscat.com`

Office document filter

The Muscat FX system can be extended to index Word, Powerpoint, Excel and other common desktop formats with the Office document filter. The Filter can be added to both the Muscat Site Indexer and Muscat Multi-Site Indexer products.

PDF filter

Muscat Limited has developed an Acrobat indexing utility which does not just index the whole PDF file, but can also optionally index page by page. When searching large document collections, Muscat can take users straight to the relevant page and highlight the search words within that page.

Customisation

Muscat Limited recognises that many companies and organisations have a variety of information held on legacy database systems, Oracle, Lotus Notes and other information systems, yet require a combined approach to information retrieval and alerting. Muscat Limited offers consultancy and customisation services to tailor indexing and search strategies to meet internal information requirements. The Muscat FX system can be adapted to meet particular customer requirements.



More product information

For more product information contact Muscat Limited at:

information@muscat.com



Index

Symbols

* 27

A

access permission 14

add

 site to multi-site index 44

advanced configuration options 45

altering priority of search term 8

AND 75

B

basic configuration 24

Basic skeleton 33

bin directory

 installed files 12

\BOOL 71

boolean query 75

BOOLQUERY tag 75

browse

 index 30

browser 44, 49

build

 index 30, 47

 multiple indexes simultaneously 39

button

 creating Next and Previous buttons 68

 displaying greyed-out buttons 69

 displaying hits pages 69

 running a query, Expand or Improve 70

C

caption

 controlling with pspec file 77

caption list 67

 inserting into query page 68



- reversing document order 78
- category 45
- category filters 71
- cgi-bin 13
 - installed files 12
 - location of scripts 13
 - Unix 16
 - Windows NT 20
- Classical skeleton 33
- combined index 61, 62
- command
 - remake-index 48
 - update-index 47
- command line interface 12
- config.txt 24, 26
 - setting Site Indexer mappings 25
- configure
 - index 35
- customise
 - score GIF 70
 - search form 63, 67

D

- database information 50
- date filters 72
 - date range by text entry 72
 - selection by year, month and day 72
 - within number of days 73
- DB file 61
- DB Size 51
 - option in browser 49
- DBS file 61, 62
- delete
 - index 50
- depth
 - of indexing 43
- directory
 - defining mappings 26
 - specifying directories to index 35
- directory mappings 24
- display
 - greyed-out buttons 69
 - search form 55
- dlist 61
- document



specifying types to include in an index 36

E

Edit sites page 44

error message 67

espec file 78

EuroFerret 75

exclude

files from index 36, 43

Expand 8, 58

hide Expand checkbox 65

inserting Expand words into query page 68

number of words in Expand list 65

expand file 67

expand_error file 67

F

file mappings 24

frame

indexing frames based sites 40

fx 5, 12

customising 63

search front end 54

fx.log 66

G

gnu tar 13

Granite skeleton 33

graphics

changing installation directory 64

installation directory (Unix) 16

installation directory (Windows NT) 21

H

height

of score GIF 64

hide

Expand checkbox 65

highlight

found words in HTML document 66

hit

specifying number displayed on query page 64



- \HITLINE 68
- \HITS 68
- hostname 26
 - enabling access to imp 13
- html
 - cgi program 12
- httpd 13

- I

- imp 5, 8, 12
 - main menu 31
 - main menu URL 23
 - restricting access 28
 - setting access permissions (Unix) 16
 - setting access permissions (Windows NT) 22
- imp.txt 24, 28
- Improve 8, 58
 - number of words added 64
- include
 - files in index 36, 43
- index
 - advanced site configuration options 45
 - browser 44
 - building 47
 - building multiple indexes simultaneously 39
 - configuring 35
 - controls 52
 - creating 30
 - deleting or renaming 50
 - depth 43
 - frames-based sites 40
 - location of index files 12
 - naming 32
 - rebuilding 50
 - searching 54
 - skeleton 33
- Index Browser 49
- index database information 50
- indexer 5, 12
 - status 50
- Indexing Controls 50
- insert current query text 68
- install script
 - running (Unix) 15
- installing Muscat FX 11



- Unix 15
- Windows NT 18
- IP address
 - setting access to imp 28

L

- Language selection 71
- language stemming
 - multi-site indexer 40
 - single-site indexer 37
- link
 - symbolic link to index (Unix) 12
- local server
 - configuring an index 35
- log
 - for search request 66
- look and feel 31
- loose proximity 65

M

- main menu 31
 - URL 23
- mappings 24
- Marble skeleton 33
- mcl (Muscat Command Line program) 12
- MDF (Muscat Document Filter) 8
- memory allocation
 - muscat-fx.cf file 62
- meta description 36, 40
- Modern skeleton 33
- msindexer 5, 12
- multi-site index
 - configuring 39
- Multi-Site Indexer
 - advanced configuration 46
- Muscat FX
 - main menu URL 23
 - pointers to installation directory 13
- muscat-fx.cf 13
 - memory allocation 62
- muscat_error file 67
- muscatfx.ini 12, 13, 33



N

- \NAME 68
 - Muscat-specific tag 67
- New Index 32
- \NEXT 68
- \NEXTOFF 68
- NOT 75

O

- OR 75

P

- page link image
 - specifying size 64
- \PAGES.x 68
- password
 - restricting imp access 28
- PDF extension
 - highlighting 66
 - page by page indexing 45
- Perl
 - system requirements 13, 18
- phrase proximity 65
- posting
 - total number in index 51
- \PREV 68
- \PREVOFF 68
- \PROB 68
- probabilistic 5
- process
 - running more than one index process 39
- proper name
 - indexing 8
- proximity 5
- proximity reordering
 - types 65
- proxy server 40
- pspec file 77

Q

- query
 - inserting text into search forms 68



- running over more than one index database 61
- submitting 56
- query expansion 8
- query file 67
- query page 55
 - number of hits 64
- query text
 - inserting into query page 68
- query.log 66

R

- real path 26
- record identity 51
- relevance 57
- relevance threshold cutoff 73
- remake-index 12, 48
- rename
 - index 50
- requirements
 - for Muscat FX installation 13
- Results page 56
 - number of hits displayed 64
- reverse
 - document order in caption list 78
- root directory
 - default mapping 27
- run
 - install script (Unix) 15
 - multiple indexes 61
 - search from different machines 26
- run_db file 61, 62

S

- \SAVE 68
- score GIF
 - setting graphics directory 64
- score GIF 57
 - customising 70
- script
 - location of 13
- \SCRIPT_NAME 69
- search
 - from static HTML page 76
 - index 54



- multiple indexes 61
- restricting by category 45
- Results page 56
- running from another machine 26
- URL 48
- using boolean filters 71
- widening 75
- Search Controls 50, 53
- search engine
 - espec file configuration 78
- search form
 - customising 63, 67
 - displaying 55
 - files governing appearance 67
- search request
 - enabling logging 66
- search techniques 5
- search term 56
- SELECT tag 71
- set access permissions 24, 28
- Setup Complete page 44
- site
 - adding to multi-site index 44
 - configuring site-specific index options 42
- site index 30
- Site Indexer
 - advanced configuration 45
 - setting up mappings 25
- Site Information 52
- size
 - of page link image 64
- skeleton 33
- static HTML form 55
 - searching from 76
- \STATS 68
- stemming 8
- superuser 14
- symbolic link 33, 36
 - to index (Unix) 12
- system requirements 13
 - access permissions 14
- T
- term
 - proximity 65



- total number in index 51
- \TERMS 68
- tight proximity 65
- top terms 74
- \TOPDOC 68
- Toptterms skeleton 33, 74
- TREATASBOOL tag 75

U

- Unix
 - installing Muscat FX 11, 15
- update-index 12, 47
- URL
 - for searching an index 48
- URL mappings 24

V

- virtual host 27
- virtual path
 - mapping between real and virtual directories 26

W

- wdf 78
- webroot 13
 - Unix 16
 - Windows NT 19
- widening a search 75
- width
 - of score GIF 64
- wildcard 27
- Windows NT
 - installing Muscat FX 11, 18
- word
 - highlighting found words 66
- word frequency 5
- word stemming 8