docoll sysadmin guide

Table of Contents

1 Introduction

This is the system administrator's guide for the docoll system.

There is an overview of the docoll system in related document "docoll system introduction".

There are separate guides for:

- docoll rsync component (for both developers and systems administrators, for server and for GNU-Linux and Windows clients).
- docoll systems administrators – for the collating and indexing component and for the interactive searching component.
- docoll collating and indexing component developers.

1.1 Related docoll documentation

In descending order of likely usefulness to a new reader:

- "docoll system introduction"
- "docoll directories and files"
- "docoll rsync server sysadmin guide"
- "docoll Windows rsync client sysadmin guide"
- "docoll GNU-Linux rsync client sysadmin guide"
- "docoll collating and indexing scripts developer's guide"
- "docoll interactive search developer's guide"
- "docoll Xapian, Omega and Apache development log"
- "docoll collating and indexing scripts development log"
- "docoll GNU-Linux rsync client development log"

2 Overview

There is an overview of the docoll system in related document "docoll system introduction". It may help to read that before reading this document.

3 System requirements

3.1 Storage space

The docoll components and are small and the Xapian database is relatively small so the storage requirements are defined by the files the rsync clients synchronise and, of those, how many are unique files of types configured to be taken into the collation.

3.2 CPU

rsync can be CPU intensive. The actual requirement depends on the number of rsync clients, the number and size of new files per synchronisation and the number of rsync clients synchronising simultaneously.

Initialising and indexing a new collation creates a heavy sustained load but the effect on the server is reduced by use of `nice` and `ionice`.

4 Software

4.1 Pre-requisites

GNU-Linux: known to work with Debian Squeeze 64-bit. Early versions known to work with CentOS 5.5 32-bit.

bash: known to work with version 4.1.5.

PostgreSQL. Known to work with 8.4.8.

rsync. Known to work with version 3.0.7.

Ruby. Known to work with version 1.9.2p180.

Webserver: known to work with Apache 2.2.16.

Xapian Core and Xapian Omega. Known to work with 1.2.8.

If docoll logs are to be mailed, an MTA such as exim, postfix or sendmail. Known to work with postscript 2.7.1.

4.2 Filters for Xapian Omega's omindex

omindex filters are explained in Appendix 1 – omindex extensions and filters.

The filters required depend on the file name extensions configured for inclusion in the collation. When you have decided what to include (default: .doc, .docx, .odp, .ods, .odt, .pdf, .pps, .ppsx, .ppt, .pptx, .rtf, .txt, .xls, .xlsx), the table in Appendix 1 – omindex extensions and filters can be consulted to find out which filters must be installed for use by omindex. The names of common packages that include these filters are listed in http://xapian.org/docs/omega/overview.html in the "omindex operation" section.

In case there is no omindex automatic filter for any of the chosen extensions or in case any of the automatic filters do not perform satisfactorily, alternative filters can be installed and docoll configured to use them:

- **Apache "Tika!"** Can filter many file types, as listed at http://tika.apache.org/0.10/formats.html. Runs around 5 times slower than omindex automatic filters and abiword but when it fails does not produce voluminous error messages as omindex automatic filters from the catdoc package have been seen to do or use huge amounts of CPU time as omindex automatic filter unrtf has been seen to do.

  It is easiest to install the complete pre-built version rather than building via Maven, for example downloading tika-app-<version>.jar from http://repo2.maven.org/maven2/org/apache/tika/tika-app/<version>.

docoll's sample configuration files assume the Tika jar is /opt/apache/tika/tika.jar, designed to be a symlink to /opt/apache/tika/tika-app-<version>.jar.

- **abiword** Can filter Word files prior to Word 2007.
- **catdoc** Can filter Word and Excel files prior to Office 2007. From http://www.wagner.pp.ru/~vitus/software/catdoc/ or from distro repositories.
- **htmltotext** If the omindex default filter for .rtf files is to be alternated with some other filter, html2text is required for the docoll script that simulates the way omindex filters .rtf files. It is available from http://www.mbayer.de/html2text or from the distro repository. More detail in sample configuration file omindex.sh.cfg, available after installing docoll.
- **unoconv** (http://dag.wieers.com/home-made/unoconv) converts files by feeding them to an OpenOffice.org server process so can convert everything that OpenOffice.org can (list at http://wiki.services.openoffice.org/wiki/Framework/Article/Filter/FilterList_OOo_3_0).

  In trials for docoll, unoconv and/or the OpenOffice.org server process were slow and not robust, probably because the OpenOffice.org python bindings used have not had a lot of test-and-fix by usage.

  docoll includes script unoconv_wrapper.sh to facilitate using unoconv. It monitors starting the OpenOffice.org server process, unoconv run time (in case the conversion has hung), unoconv error messages (one does not indicate conversion failure) and checks for empty output. Occasionally unoconv_wrapper.sh has been seen to become a zombie on termination; when this happens, omindex cleans up after 300 seconds.

  According to the unoconv source code repository, unoconv was updated for LibreOffice in October 2011. At the time of writing, this version has not been tested with docoll.
- **PyODConverter and JODConverter** (http://www.artofsolving.com/opensource) work with an OpenOffice.org or LibreOffice server process, similar to unoconv. At the time of writing, they have not been tried with docoll.

None of the tested filters have been found 100% effective. To help with this issue, docoll can be configured with multiple filters for each extension. Each time docoll starts indexing, if it has been configured with multiple filters for an extension, it will choose one at random. In this way, over multiple index runs, if a previous filter has failed to convert a file to text, every filter will be used on it.

4.3 docoll

docoll server software, including sample configuration files and omega templates, is in a compressed tar archive called docoll_server-<version>.tgz.

> Note: docoll software for rsync clients is documented in related documents "docoll GNU-Linux rsync client sysadmin guide" and "docoll Windows rsync client sysadmin guide".

5 docoll setup

5.1 Pre-requisites

Install any of the pre-requisite software listed above that is not already installed.

5.1.1 Apache

The only Apache module used by docoll that is not usually enabled by default is "rewrite". Ensure it is available by running, as root:

```
a2enmod rewrite
```

If the rewrite module was not already enabled, the command will advise you to restart Apache. It need not be done because the docoll installation procedure includes making Apache re-load its configuration.

5.1.2 Apache and Xapian Omega

docoll assumes the Apache configuration has `ScriptAlias /cgi-bin/ /usr/lib/cgi-bin/` and Xapian Omega's omega CGI executable is /usr/lib/cgi-bin/omega/omega. If not, adjustments will be required to one or more of:

- The Apache configuration.
- The Xapian Omega installation.
- /etc/apache2/conf.d/docoll (after it has been created during docoll setup).

5.1.3 PostgreSQL

As the PostgreSQL user (normally postgres), create the docoll user and make a note of the password set:

```
createuser --pwprompt docoll
Enter password for new role:
Enter it again:
Shall the new role be a superuser? (y/n) n
Shall the new role be allowed to create databases? (y/n) y
Shall the new role be allowed to create more new roles? (y/n) n
```

5.1.4 Tika! (Apache Tika!)

The sample docoll configuration file omindex.sh.cfg assumes that the Tika jar is available as /opt/apache/tika/tika.jar.

If the Tika jar is /opt/apache/tika/tika-app-0.9.jar:

```
cd /opt/apache/tika && ln -s tika-app-0.9.jar tika.jar
```

5.1.5 Xapian Core and Xapian Omega

In case no packages of the desired versions are available, commands to build and install from source are given in Appendix 2 – Installing Xapian 1.2.8 from source.

5.2 Synchronisation (rsync) setup

> Note: if you do not want to set up file synchronisation from rsync clients now, you can skip this stage and create a docoll source files tree for testing. The procedure to do so is described below.

The rsync server setup procedure is described in related document "docoll rsync server sysadmin guide".

The rsync client setup procedures are described in related documents "docoll GNU-Linux rsync client sysadmin guide" and "docoll Windows rsync client sysadmin guide".

5.3 docoll server software installation

For illustration, the procedure to install version 0.7.3 is described.

> Note: the directory structure is the one described in related document "docoll directories and files". In case you want to change it, docoll is designed to be configurable for any directory layout.

As root …

1. Create group docoll and user docoll.
2. Create directories:

```
dirs='
    /etc/opt/docoll
    /opt/docoll
    /srv/docoll
    /srv/rsync/docoll
    /var/opt/docoll
    /var/log/docoll
    /var/www/docoll
'
mkdir -p $dirs
chown docoll:docoll $dirs
```

As user docoll …

1. Install the docoll server software:

```
version=0.7.3
tar -xvzf docoll_server-$version.tgz --directory /
```

As root …

1. Install the docoll Apache configuration file and load it:

```
version=0.7.3
cp -p /opt/docoll/$version/samples/apache /etc/apache2/conf.d/docoll
chown root:root /etc/apache2/conf.d/docoll
apachectl -k graceful
```

5.4 docoll default instance setup

Notes:

Multiple instances of docoll can be run on a server. Each instance may be set up to have its own:

- Sources – one or more trees of files to collate
- Collation – files copied from the sources with duplicates removed
- Collating and indexing (C+I) scripts version
- Collating and indexing (C+I) scripts configuration
- omega CGI executable version
- omega CGI executable templates
- Log files

The directory structure is the one described in related document "docoll directories and files". In case you want to change it, docoll is designed to be configurable for any directory layout.

For illustration, the procedure to set up instance "default" is described.

As user docoll …

1. Link the program directory for the instance:

```
version=0.7.3; instance=default
ln -s /opt/docoll/$version/bin /opt/docoll/$instance
```

2. Create directories and symlinks for the instance:

```
instance=default
mkdir -p \
    /etc/opt/docoll/$instance \
    /srv/docoll/$instance \
    /srv/rsync/docoll/$instance \
    /var/log/docoll/$instance \
    /var/opt/docoll/$instance/backup \
    /var/opt/docoll/$instance/sources \
    /var/opt/docoll/$instance/xapian_db \
    /var/www/docoll/$instance
ln -s /srv/docoll/$instance /var/www/docoll/$instance/hits
cd /var/opt/docoll/$instance/xapian_db && ln -s . $instance
```

3. Create the collating scripts' database for the default instance:

```
instance=default
createdb --owner docoll docoll_$instance
```

4. Create and customise the configuration files for the instance:

```
version=0.7.3; instance=default
cd /opt/docoll/$version/samples \
    && cp -p -R *.cfg omega.conf templates /etc/opt/docoll/$instance \
    && cd /etc/opt/docoll/$instance \
    && sed -i "s/%instance%/$instance/" \
        collate.cfg omega.conf omindex.sh.cfg run_scripts.cfg templates/inc/instance_cfg
```

5. Optionally, docoll collation source file tree(s) for the instance can be created under `/var/opt/docoll/$instance/sources`
6. Adjust the instance's configuration files:

   The only necessary change is to set the PostgreSQL user docoll's password in collate.cfg.

   The instance's configuration files are in /etc/opt/docoll/$instance.

   Changes you may like to consider:

   collate.cfg

   Custom docoll source directories can be set in the SourceRootDirs section, in which case the LeadingDirsToStrip section should be changed to match.

   docoll's default extensions list (.doc, .docx, .odp, .ods, .odt, .pdf, .pps, .ppsx, .ppt, .pptx, .rtf, .txt, .xls, .xlsx) can be changed.

   omindex.sh.cfg The default filters used by omindex can be changed. They are omindex's defaults and are faster than many of the alternatives so it is better to defer changing them until after the first run.

   run_scripts.cfg Log emailing can be set up. For testing, this can be set to "always".

   templates/inc/instance_cfg The search web page title can be changed from "docoll <instance> instance search" in.
7. Set up a cron job to run docoll's collating and indexing (C+I) scripts:

   The first run takes longest so you may like to run the C+I scripts manually a few times before setting up the cron job to find out how long the C+I scripts normally take to run. The command to do so are described in 6.2 Running the collating and indexing (C+I) scripts.

   The choice of frequency depends on:

   - How quickly the files in the docoll sources change.
   - How quickly you want new documents to appear in the interactive search results.
   - How long the C+I scripts take to run.

   The choice of time-of-day is not critical because:

   - After the first run, to create the Xapian database, interactive searching is available while the scripts are running.
   - The default configuration for the scripts sets `nice` and `ionice` values for low impact on server performance.

   There is a sample cron job for the default instance. It can be installed and then modified to suit your frequency and time of day requirements. As user docoll ...

```
version=0.7.3
crontab /opt/docoll/$version/samples/cronjob
crontab -e
```

6 Initial run and test

6.1 Populating the docoll source file tree(s)

If rsync clients have been set up, allow time for them to populate the docoll sources or run the synchronisation manually as described in related documents "docoll GNU-Linux rsync client sysadmin guide" and "docoll Windows rsync client sysadmin guide". These guides, along with related document "docoll rsync server sysadmin guide", also describe how to check the logs.

If rsync clients have not been set up, you must have populated /var/opt/docoll/default/sources.

6.2 Running the collating and indexing (C+I) scripts

If you have created a cron job to run the C+I scripts, allow time for the docoll scripts to populate the docoll collation.

Alternatively, the C+I scripts can be run manually. If there are many files in the sources, the scripts will write voluminous logging to the terminal. This can be suppressed by using SET_HAVE_TTY_FALSE as shown below:

```
export SET_HAVE_TTY_FALSE=true    # Optional, to force logging to log files
instance=default
cd /opt/docoll/$instance \
    && nohup ./run_scripts.sh /etc/opt/docoll/$instance/run_scripts.cfg &
```

The log files are written in /var/log/docoll/default. When the run has finished, the run_scripts.sh log will be the latest one. In case an omindex filter hangs, `top` will not show any processor usage by docoll processes until omindex kills it after 300 seconds.

Aliases can be useful, for example:

```
alias dbin='cd /opt/docoll/default && lrt'
alias dcfg='cd /etc/opt/docoll/default && lrt'
alias dlog='cd /var/log/docoll/default && lrt | tail -8 && du -hs .'
alias dtbin='cd /opt/docoll/test && lrt'
alias dtcfg='cd /etc/opt/docoll/test && lrt'
alias dtlog='cd /var/log/docoll/test && lrt | tail -8 && du -hs .'
```

6.3 Test

Browse http://<webserver name or IP address>/docoll, where the part in angled brackets is substituted by appropriate string, do a search and ensure that a file in the search hit list can be viewed or downloaded.

7 Monitoring

/var/log/rsyncd.log provides useful information, not only about errors but, by showing which files are being transferred, about over-inclusive or under-inclusive configuration of the rsync clients.

The run_scripts.sh log in /var/log/docoll/default provides a summary. It shows whether the scripts it called reported errors, in which case their logs can be viewed for more detail. It also shows the run times of the scripts it called; if these are unusually short or long, it probably indicates trouble.

If run_scripts.cfg has been edited to email the run_scripts.sh log, the log will be emailed according to the always|warning|error setting.

8 Setting up non-default instances

The procedure to set up a non-default instance is as described in 5.4 docoll default instance setup except:

- In the `instance=default` commands, "default" is changed to the new instance name.
- If a cron job is required for the new instance (the source files will change), docoll's crontab must be changed to include a new line for each instance, changing all three "default" strings to the new instance name:

  ```
  /var/opt/docoll/default && /opt/docoll/default/run_scripts.sh
  /etc/opt/docoll/default/run_scripts.cfg
  ```

- The interactive search link for the new instance is http://<webserver name or IP address>/docoll/<instance name> where the parts in angled brackets are substituted by appropriate strings.

A further instance cron job may be added by copying the default line and changing all three "default" strings to the further instance name.

9 Miscellaneous techniquies

9.1 Switching collations

Xapian's omindex records paths relative to its directory argument. This allows a new collation to be indexed and then switched into production. For example, if the production collation is /srv/docoll/default and the production Xapian database is /var/opt/docoll/default/xapian_db, it is possible to index a new collation at /srv/docoll/new into a Xapian database in /var/opt/docoll/new/xapian_db and then put the new version into production by:

```
rm  /var/opt/docoll/default/xapian_db/*
cp -p /var/opt/docoll/new/xapian_db/* /var/opt/docoll/default/xapian_db
mv /srv/docoll/default /srv/docoll/default.old
mv /srv/docoll/new /srv/docoll/default
```

9.2 Restoring a collating scripts' database

```
instance=default
export PGPASSWORD=<password>
pg_restore \
  --dbname=docoll \
  --host=localhost \
  --port=5432 \
  --verbose \
  -U docoll \
  /var/opt/docoll/$instance/backup/docoll_$instance.<timestamp>
```

9.3 Restoring a Xapian index database

```
instance=default
```

```
cd /var/opt/docoll/$instance/ \
    && rm -f xapian_db/* \
    && tar xzvf backup/Xapian_index.<timestamp>.tgz -d xapian_db
```

10 Problem investigation

Apart from the usual sysadmin issues such as running out of disk space and misconfiguration, most docoll problems have been docoll scripts bugs. These are developer issues; the related documents for developers may help investigate them: "docoll collating and indexing scripts developer's guide" and "docoll interactive search developer's guide".

11 Appendix 1 – omindex extensions and filters

Omindex:

- Has built-in support for indexing HTML, PHP, text files, CSV (Comma-Separated Values) files, and AbiWord documents.
- Automatically uses some programs as filters if they are installed. Filters are programs that convert files to plain text for omindex to index.
- Accepts an option to specify the command to be used to filter specific "MIME types". Up to at least version 1.2.7, these are not true "MIME types"; omindex derives them from the file name extension.

  Note: up to at least Xapian Omega 1.2.7, these filter command lines must accept the name of the input file as their last argument.

Extension	"MIME type"	Type	Built-in	Automatic	Notes or filter for Automatic
abw	application/x-abiword	AbiWord	Y	N/A	
ai	application/postscript			Y	Requires both ps2pdf and pdftotext
csv	text/csv	CSV	Y	N/A	
deb	application/x-debian-package			Y	dpkg-deb
djv	image/vnd.djvu			Y	djvutxt
djvu	image/vnd.djvu			Y	djvutxt
doc	application/msword			Y	antiword
docm	application/vnd.openxmlformats-officedocument.wordprocessingml.document				
docx	application/vnd.openxmlformats-officedocument.wordprocessingml.document	Word 2007			
dot	application/msword	Word template		Y	antiword
dotm	application/vnd.openxmlformats-officedocument.wordprocessingml.template				
dotx	application/vnd.openxmlformats-officedocument.wordprocessingml.template	Word 2007 template			
dvi	application/x-dvi			Y	catdvi
eps	application/postscript			Y	Requires both ps2pdf and pdftotext

Extension	"MIME type"	Type	Built-in	Automatic	Notes or filter for Automatic
htm	text/html	HTML	Y	N/A	
html	text/html	HTML	Y	N/A	
msg	application/vnd.ms-outlook	Outlook .msg email		Y	perl with Email::Outlook::Message and HTML::Parser modules
odb	application/vnd.oasis.opendocument.database			Y	unzip
odc	application/vnd.oasis.opendocument.chart			Y	unzip
odf	application/vnd.oasis.opendocument.formula			Y	unzip
odg	application/vnd.oasis.opendocument.graphics			Y	unzip
odi	application/vnd.oasis.opendocument.image			Y	unzip
odm	application/vnd.oasis.opendocument.text-master			Y	unzip
odp	application/vnd.oasis.opendocument.presentation			Y	unzip
ods	application/vnd.oasis.opendocument.spreadsheet			Y	unzip
odt	application/vnd.oasis.opendocument.text			Y	unzip
otc	application/vnd.oasis.opendocument.chart-template			Y	unzip
otf	application/vnd.oasis.opendocument.formula-template			Y	unzip
otg	application/vnd.oasis.opendocument.graphics-template			Y	unzip
oth	application/vnd.oasis.opendocument.text-web			Y	unzip
oti	application/vnd.oasis.opendocument.image-template			Y	unzip
otp	application/vnd.oasis.opendocument.presentation-template			Y	unzip
ots	application/vnd.oasis.opendocument.spreadsheet-template			Y	unzip
ott	application/vnd.oasis.opendocument.text-template			Y	unzip
pdf	application/pdf			Y	pdftotext
php	text/html	HTML	Y	N/A	The omindex HTML parser ignores PHP code.
pl	text/x-perl	Text	Y	N/A	pod2text
pm	text/x-perl	Text	Y	N/A	pod2text
pod	text/x-perl	Text	Y	N/A	pod2text
potm	application/vnd.openxmlformats-officedocument.presentationml.template				
potx	application/vnd.openxmlformats-officedocument.presentationml.template	PowerPoint 2007 template			
pps	application/vnd.ms-powerpoint	Powerpoint slideshow			
ppsm	application/vnd.openxmlformats-officedocument.presentationml.slideshow				
ppsx	application/vnd.openxmlformats-officedocument.presentationml.slideshow	PowerPoint 2007 slideshow			
ppt	application/vnd.ms-powerpoint			Y	catppt
pptm	application/vnd.openxmlformats-officedocument.presentationml.presentation				
pptx	application/vnd.openxmlformats-officedocument.presentationml.presentation	PowerPoint 2007 presentation			
ps	application/postscript			Y	Requires both ps2pdf and pdftotext
rpm	application/x-redhat-package-manager			Y	rpm
rtf	text/rtf			Y	unrtf
shtml	text/html	HTML	Y	N/A	
stc	application/vnd.sun.xml.calc.template			Y	unzip

Extension	"MIME type"	Type	Built-in	Automatic	Notes or filter for Automatic
std	application/vnd.sun.xml.draw.template			Y	unzip
sti	application/vnd.sun.xml.impress.template			Y	unzip
stw	application/vnd.sun.xml.writer.template			Y	unzip
svg	image/svg+xml				
sxc	application/vnd.sun.xml.calc			Y	unzip
sxd	application/vnd.sun.xml.draw			Y	unzip
sxg	application/vnd.sun.xml.writer.global			Y	unzip
sxi	application/vnd.sun.xml.impress			Y	unzip
sxm	application/vnd.sun.xml.math			Y	unzip
sxw	application/vnd.sun.xml.writer			Y	unzip
text	text/plain	Text	Y	N/A	
txt	text/plain	Text	Y	N/A	
udeb	application/x-debian-package			Y	dpkg-deb
wpd	application/vnd.wordperfect			Y	wpd2text
wps	application/vnd.ms-works			Y	wps2text
wpt	application/vnd.ms-works	Works template		Y	wps2text
xlb	application/vnd.ms-excel			Y	xls2csv
xlr	application/vnd.ms-excel	Later Microsoft Works		Y	xls2csv
xls	application/vnd.ms-excel			Y	xls2csv
xlsm	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet				
xlsx	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	Excel 2007			
xlt	application/vnd.ms-excel	Excel template		Y	xls2csv
xltm	application/vnd.openxmlformats-officedocument.spreadsheetml.template				
xltx	application/vnd.openxmlformats-officedocument.spreadsheetml.template	Excel 2007 template			
xps	application/vnd.ms-xpsdocument			Y	unzip
zabw	application/x-abiword-compressed	AbiWord compressed	Y	N/A	Also requires gzip

Notes on sources of information for the table:

1. General information for the table came from the "omindex operation" section of http://xapian.org/docs/omega/overview.html.
2. File name extension to "MIME type" mappings came from the omindex source code. The latest version can be found at https://gitorious.org/xapian/xapian/blobs/master/xapian-applications/omega/omindex.cc in the mime_map definition which started on line 1026 at the time of writing.

   The command used to prepare data copied from that page was:

   ```
   grep -E -v '^([0-9])*[[:space:]]*$|^[[:space:]]*//' | sed -e 's/
   [[:space:]]*mime_map\["//' -e 's/"] = "/\t/' -e 's/"; \/\/ /\t/' -e 's/";$//'
   ```

   The tabs in the output allowed it to conveniently be converted into a table in OpenOffice Writer and then copied and pasted into the table above.

12 Appendix 2 – Installing Xapian 1.2.8 from source

In case no packages of the desired versions are available, here are commands and a script to build and install from source

```
ver=1.2.8
repository=/root/Repository
prefix=/usr

pkg=xapian-core-$ver
cd $repository \
    && wget http://oligarchy.co.uk/xapian/$ver/xapian-core-$ver.tar.gz
cp $repository/$pkg.tar.gz /tmp \
    && cd /tmp \
    && tar xzvf $pkg.tar.gz \
    && cd $pkg \
    && ./configure --prefix=/usr \
    && make \
    && make install

pkg=xapian-omega-$ver
cd $repository \
    && wget http://oligarchy.co.uk/xapian/$ver/xapian-omega-$ver.tar.gz
cp /root/Repository/$pkg.tar.gz /tmp \
    && cd /tmp \
    && tar xzvf $pkg.tar.gz \
    && cd $pkg \
    && ./configure --prefix=/usr \
    && make
```

The default Xapian Omega source installation does not suit docoll's requirements which are based on how Debian packages install it. This quick-and-dirty script, specifically for Xapian Omega 1.2.8, installs it as required for docoll.

It normally installs to the default locations but can be made to install everything to another directory, given as an argument. This is useful if you want to generate a list of the files installed for use when removing it.

```
#! /bin/bash

# Installs Xapian Omega 1.2.8 for docoll.
# Inspired by the way Olly Betts (Xapian developer) packages it for Debian.

ver=1.2.8

if [[ $1 != "" ]]; then
    root=$1
    if [[ $root =~ ^/ ]]; then
        echo "Installing version $ver to $root"
    else
        echo "Installation directory must begin with /"
        exit 1
    fi
else
    read -p 'Install for real? (Y to confirm): '
    [[ $REPLY != Y ]] && exit 0
    root=
fi

dir=/tmp/xapian-omega-$ver
cd $dir
if [[ $? -ne 0 ]]; then
```

```
    echo "The source code must have been built in $dir before running this script"
    exit 1
fi

#  Create required directories if not installing to /
if [[ ! -d $root ]];then
    mkdir -p \
        $root/etc \
        $root/usr/bin \
        $root/usr/lib \
        $root/usr/share/doc \
        $root/usr/share/images \
        $root/usr/share/man/man1
fi

chown -R root:root *

# Install as Olly's pro-forma file list as far as practicable
cp -p omega.conf $root/etc/omega.conf
cp -p omindex $root/usr/bin/omindex
cp -p scriptindex $root/usr/bin/scriptindex
mkdir -p $root/usr/lib/cgi-bin/omega/
cp -p omega $root/usr/lib/cgi-bin/omega/omega
mkdir -p $root/usr/lib/xapian-omega/bin/
cp -p outlookmsg2html $root/usr/lib/xapian-omega/bin/outlookmsg2html
#/usr/share/doc-base/xapian-omega-docs   Not found
mkdir -p $root/usr/share/doc/xapian-omega/examples/
#/usr/share/doc/xapian-omega/TODO.Debian   Not found
#/usr/share/doc/xapian-omega/changelog.Debian.gz   Not found
#/usr/share/doc/xapian-omega/changelog.gz   Not found
#/usr/share/doc/xapian-omega/copyright   Not found
cp -p dbi2omega $root/usr/share/doc/xapian-omega/examples/dbi2omega
cp -p htdig2omega $root/usr/share/doc/xapian-omega/examples/htdig2omega
cp -p htdig2omega.script $root/usr/share/doc/xapian-omega/examples/htdig2omega.script
cp -p mbox2omega $root/usr/share/doc/xapian-omega/examples/mbox2omega
cp -p mbox2omega.script $root/usr/share/doc/xapian-omega/examples/mbox2omega.script
mkdir -p $root/usr/share/xapian-omega/templates/
ln -s $root/usr/share/xapian-omega/templates/ $root/usr/share/doc/xapian-
omega/examples/templates
cp -p docs/* $root/usr/share/doc/xapian-omega/
rm $root/usr/share/doc/xapian-omega/Makefile*
mkdir -p $root/usr/share/images/xapian-omega/
cp -p images/* $root/usr/share/images/xapian-omega/
(
    cp -p omindex.1 scriptindex.1 /tmp \
        && cd /tmp \
        && gzip -9 omindex.1 scriptindex.1 \
        && cp -p omindex.1.gz scriptindex.1.gz $root/usr/share/man/man1/
    rm -f omindex.1.gz scriptindex.1.gz
)
mkdir -p $root/usr/share/xapian-omega/templates/inc
rsync -a --quiet templates/ $root/usr/share/xapian-omega/templates/

# Extra to Olly's proforma file list
cp -p AUTHORS COPYING ChangeLog NEWS README TODO $root/usr/share/doc/xapian-omega/ #
Equivalent to some of the "Not found"?
#  extra/omegascript.vim    Not found
if [[ ! -d $root/var/lib/omega/templates ]]; then
    mkdir -p $root/var/lib/omega/templates
    rsync -a --quiet templates/ $root/var/lib/omega/templates
else
   echo "$root/var/lib/omega/templates already exists; not creating or populating it"
   echo "The $ver example templates are installed in $root/usr/share/xapian-
```

```
omega/templates/"
fi
mkdir -p $root/var/log/omega
mkdir -p $root/var/lib/omega/cdb
```

13 Appendix 3 – GNU Free Documentation License

Version 1.3, 3 November 2008

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position

regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "publisher" means any person or entity that distributes copies of the Document to the public.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

* A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

* B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
* C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
* D. Preserve all the copyright notices of the Document.
* E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
* F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
* G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
* H. Include an unaltered copy of this License.
* I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
* J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
* K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
* L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
* M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
* N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
* O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the

previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.