docoll collating and indexing scripts developer's guide

Table of Contents

1 Introduction

This is the developer's guide for the collating and indexing component of the docoll system.

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

1.1 Related docoll documentation

In descending order of likely usefulness to a new reader:

- "docoll system introduction"
- "docoll directories and files"
- "docoll server sysadmin guide" (how it all fits together; detailed info on omindex filters)
- "docoll Ruby coding standards"
- "docoll packager's guide" (also useful for making backups of work in progress)
- RELEASE (needs updating with changes made)

1.2 Xapian resources

- Xapian home page: http://xapian.org/
- Xapian documentation: http://xapian.org/docs/
- Xapian omega overview (includes omindex): http://xapian.org/docs/omega/overview.html
- Xapian WIKI: http://trac.xapian.org/wiki
- Xapian user mailing list archives: http://lists.xapian.org/pipermail/xapian-discuss/ (search via http://www.google.co.in/advanced_search?hl=en&num=30&lr=lang_en&ft=i&as_sitesearch=lists.xapian.org%2Fpipermail%2Fxapian-discuss&as_qdr=all&as_occt=any)
- Xapian devel mailing list archives: http://lists.xapian.org/pipermail/xapian-devel/ (search via http://www.google.co.in/advanced_search?hl=en&num=30&lr=lang_en&ft=i&as_sitesearch=lists.xapian.org/pipermail/xapian-devel&as_qdr=all&as_occt=any)

2 Data files' directories

General information about directories is in related document "docoll directories and files".

2.1 Sources

As configured by the SourceRootDirs parameter. There may be several sources, for example /srv/rsync/docoll/<instance> and /var/opt/docoll/<instance>/sources.

2.2 Collation

As configured by the CollationRootDir parameter. There can be only one, for example /srv/docoll/<instance>.

3 Databases

There are two databases in a docoll system.

3.1 PostgreSQL

This database has three purposes:

1. To accelerate the process of "finger printing" files by recording their expensive-to-generate checksums (md5 and sha1). As the scripts find source files suitable for the collation, they lookup the file in the database by path and check the mtime. If the mtime has not changed there is no need to re-compute the md5 and sha1 sums.
2. To identify any file/inode in the collation which is identical (same checksums) to a source file.
3. To record the paths for each file/inode in the collation.

There are three tables/relations in the docoll Postgres database:

```
            List of relations
 Schema |      Name      | Type  | Owner
--------+----------------+-------+--------
 public | collated_files | table | docoll
 public | collated_paths | table | docoll
 public | source_files   | table | docoll
(3 rows)

             Table "public.collated_files"
  Column   |            Type             | Modifiers
-----------+-----------------------------+-----------
 inode     | integer                     | not null
 md5       | character(32)               | not null
 mime_type | text                        | not null
 mtime     | timestamp without time zone | not null
 sha1      | character(40)               | not null
Indexes:
    "collated_files_md5_key" UNIQUE, btree (md5, sha1)

Table "public.collated_paths"
 Column |  Type   | Modifiers
--------+---------+-----------
 inode  | integer | not null
 path   | text    |
Indexes:
    "collated_paths_path_key" UNIQUE, btree (path)

              Table "public.source_files"
  Column   |            Type             | Modifiers
-----------+-----------------------------+-----------
 inode     | integer                     | not null
 md5       | character(32)               | not null
 mime_type | text                        | not null
 mtime     | timestamp without time zone | not null
 path      | text                        |
 sha1      | character(40)               | not null
Indexes:
    "source_files_path_key" UNIQUE, btree (path)
```

3.2 Xapian

The Xapian database is generated by Xapian's omindex utility and used by Xapian's omega CGI script to service interactive text searches of the documents in the collation.

The Xapian database is in directory /var/opt/docoll/<instance>/xapian_db/. This is configured by /etc/opt/docoll/<instance>/omega.conf on the database_dir line. Omega looks for the database in a subdirectory of database_dir, named on the CGI query string's DB parameter. This is serviced by a symlink in /var/opt/docoll/<instance>/xapian_db/ called <instance> and pointing to its containing directory:

```
docoll@CW8vDS:/var/opt/docoll/default/xapian_db$ ll default
```

```
lrwxrwxrwx 1 docoll docoll 1 Dec 10 19:32 default -> .
```

4 Scripts

The scripts are installed in /opt/docoll/<version>/bin, symbolically linked as /opt/docoll/<instance>. For example, on a production system runing docoll 0.7.3, during 0.7.6 beta testing:

```
docoll@LS1:~$ ls -l /opt/docoll
total 0
drwxr-xr-x 4 docoll docoll 40 Dec 12 17:47 0.7.3
drwxr-xr-x 4 docoll docoll 40 Mar 24 13:24 0.7.6.beta.4
lrwxrwxrwx 1 docoll docoll 21 Dec 13 22:26 default -> /opt/docoll/0.7.3/bin
lrwxrwxrwx 1 docoll docoll 21 Mar 30 22:42 test -> /opt/docoll/0.7.6/bin
```

They are written in bash (*.sh) and Ruby (*.rb):

```
docoll@CW8vDS:~$ ls -l /opt/docoll/0.7.6/bin/*[bh]
-r-xr-xr-x 1 docoll docoll 2.5K Oct 27 12:43 /opt/docoll/0.7.6/bin/CollatedFile.rb
-r-xr-xr-x 1 docoll docoll 2.5K Mar 26 18:35 /opt/docoll/0.7.6/bin/Log.rb
-r-xr-xr-x 1 docoll docoll 3.1K Oct  9 12:03 /opt/docoll/0.7.6/bin/SourceFile.rb
-r-xr-xr-x 1 docoll docoll  11K Mar 28 14:33 /opt/docoll/0.7.6/bin/analyse_omindex_log.sh
-r-xr-xr-x 1 docoll docoll  12K Mar 26 15:55 /opt/docoll/0.7.6/bin/backup_db.sh
-r--r--r-- 1 docoll docoll  13K Mar 23 14:40 /opt/docoll/0.7.6/bin/bash_lib.sh
-r-xr-xr-x 1 docoll docoll  15K Mar 30 21:57 /opt/docoll/0.7.6/bin/clean.rb
-r-xr-xr-x 1 docoll docoll  16K Mar 30 22:07 /opt/docoll/0.7.6/bin/clean_omindex_log.rb
-r-xr-xr-x 1 docoll docoll  16K Mar 30 21:29 /opt/docoll/0.7.6/bin/collate.rb
-r-xr-xr-x 1 docoll docoll  17K Mar 28 14:27 /opt/docoll/0.7.6/bin/omindex.sh
-r-xr-xr-x 1 docoll docoll 7.0K Mar 28 10:36 /opt/docoll/0.7.6/bin/parse_cfg_for_bash.rb
-r-xr-xr-x 1 docoll docoll 1.1K Oct 28 16:05 /opt/docoll/0.7.6/bin/pdftotext_wrapper.sh
-r--r--r-- 1 docoll docoll  18K Mar 30 21:53 /opt/docoll/0.7.6/bin/ruby_db_lib.rb
-r--r--r-- 1 docoll docoll  14K Mar 24 16:05 /opt/docoll/0.7.6/bin/ruby_lib.rb
-r-xr-xr-x 1 docoll docoll  17K Mar 29 22:26 /opt/docoll/0.7.6/bin/run_scripts.sh
-r-xr-xr-x 1 docoll docoll  12K Dec 15 01:09 /opt/docoll/0.7.6/bin/unoconv_wrapper.sh
-r-xr-xr-x 1 docoll docoll 1.2K Dec  8 07:12 /opt/docoll/0.7.6/bin/unrtf_wrapper.sh
```

4.1 Libraries

bash_lib.sh is the bash library.

ruby_db_lib.rb is the Ruby database library (facilitates changing from Postgres to another database).

ruby_lib.rb is the Ruby general library.

4.2 Ruby file naming and coding standards

The format of the *.rb file names is explained in related document "docoll Ruby coding standards".

4.3 Ruby configuration file parsing scripts

Configuration files for Ruby scripts are parsed by Ruby library docoll_lib.rb, specifically by ParseConfigFile.

Ruby script parse_cfg_for_bash.rb allows bash scripts to parse the same config files.

Historical note

Existing Ruby configuration file parsing utilities were studied:

- YAML (example at http://stackoverflow.com/questions/4375530/ruby-configuration-file-parser-combined-with-optionparser).
- drks' ParseConfig (http://www.5dollarwhitebox.org/drupal/?q=node/21).
- Hans' parse_config (http://otype.de/index.php?id=151).

None of the above gave both an easily legible configuration file syntax and the power to load hashes as well as arrays.

4.4 Log.rb – Ruby logging

Historical note

Existing utility Logger was examined but found wanting. Log.rb was written with the required functionality.

4.5 run_scripts.sh

run_scripts.sh is the master script that runs all the others. It is intended to be run as a cron job but may be run manually, for example when testing.

run_scripts.sh' log contains sufficient excerpts of logs from the scripts it runs to be used to determine whether individual logs need to be examined.

The remaining scripts are described in the order that run_scripts.sh runs them.

4.6 collate.rb

1. Initialise: initialise configuration parameters, parse any configuration file, parse remaining command line items, set up logging, set signal traps, connect to database and initialise database if empty.
2. For each file in the configured sources …
   a) Iterate loop if file is excluded for any reason (directory, symlink, not a file, too small, not an included extension, an excluded MIME type or extension).

      TODO: remove exclusion by MIME type or extension; add inclusion by MIME type.

   b) Instantiate a SourceFile object from the path and file (path, inode, mtime). Lookup path in database. Delete from database if path's file has changed. If path is in database, get checksums (md5 and sha1) and MIME type from database, else get from the file and write to database.

   c) Iterate loop if MIME type or extension are excluded. Note: this is redundant because done already.

   d) Strip any leading directory configured in LeadingDirsToStrip. This normally includes the source top level directory.

   e) Determine the "equivalent path" by prefixing the path with CollationRootDir.

   f) Instantiate a CollatedFile object. Look in the database for a collated file with same checksums (md5 and sha1) as the SourceFile. If found, get inode, MIME type and mtime from the database. Otherwise get them from the source file.

g) If there is no identical file in the collation, copy the source file into the collation at the equivalent path. Otherwise, if the equivalent path does not exist, create it by hard linking.

h) Write any changes in the collation (new file or new path) to the database.

i) If the source file has an earlier mtime than the collated file, set the collated file's mtime to the source file's and update the database. Note: this is required because interactive users can search for files by date. In case the sources contain duplicates of a file with a variety of mtimes (the result of copying files without preserving the mtimes) the best we can do is set the collated file to the earliest.

3. Delete old log files.
4. Finalise: close the database connection, write summary messages to the log.

4.7 backup_db.sh

1. Initialise: source the bash library (includes setting signal traps), parse command line, set up redirection and logging, parse the configuration file (using parse_cfg_for_bash.rb). The configuration file is normally the same one used by collate.rb.
2. Use PostgreSQL utility pg_dump to dump the PostgreSQL database to the configured directory, in the production system /var/opt/docoll/backup.
3. Delete old database dump files.
4. Finalise: write summary messages to the log.

4.8 omindex.sh

1. Initialise: source the bash library (includes setting signal traps), parse command line, set up redirection and logging, parse the configuration file (using parse_cfg_for_bash.rb). The configuration file is normally the same one used by collate.rb.
2. Use Xapian utility omindex to index the collation. TODO: make filters configurable.
3. Analyse the failure rates for each file type (using analyse_omindex_failure_rates.sh).
4. Finalise: write summary messages to the log.

4.8.1 omindex filters

The choice of filter executables for use with the omindex command is critical. The script analyse_omindex_failure_rates.sh was written to help choose the best filters.

4.8.2 clean_omindex_log.rb

Run by omindex.sh when configured with `omindex log cleaning = true` (default).

Cleans the omindex error log of messages that are not useful to the docoll systems administrator.

4.8.3 analyse_omindex_failure_rates.sh

Run by omindex.sh. Produces log output like:

```
06:18:17 analyse_omindex_log.sh:
   doc files: in tree: 20415, failed: 138   .67%
  docx files: in tree:   899, failed:   0
   odp files: in tree:     7, failed:   0
   ods files: in tree:    98, failed:   0
   odt files: in tree:   171, failed:   0
```

```
   pdf files: in tree: 12121, failed:   0
   pps files: in tree:    90, failed:   0
  ppsx files: in tree:    14, failed:   0
   ppt files: in tree:   579, failed:   7  1.20%
  pptx files: in tree:    55, failed:   0
   rtf files: in tree:  2031, failed:   1   .04%
   txt files: in tree:  6370, failed:   0
   xls files: in tree:  6621, failed:  13   .19%
  xlsx files: in tree:   302, failed:   0
  Total files in tree: 49773
06:18:17 WARN: 309 files skipped by omindex. List appended to
/var/log/docoll/test/omindex_error.12-03-31@00:54.log
```

4.9 clean.rb

1. Initialise: initialise configuration parameters, parse any configuration file, parse remaining command line items, set up logging, set signal traps, connect to database and initialise database if empty.
2. For each path in the file system under the collation directory …
   a) If the path is an empty directory, remove it.
   b) Iterate the loop if the path is not for a file (as defined by Ruby's File.file method) or a symlink.
   c) If the path is excluded for any reason (its file is too small, a symlink, not an included extension, an excluded MIME type or extension) remove it and iterate the loop. This would be effective if the configuration's minimum file size, included extensions or excluded types or extensions were changed.
   d) If the path is not already in the database:
      i. If the database has a collated file with same checksums (md5 and sha1), get its inode otherwise insert the collated file in the database using the path to get its inode and other properties.
      ii. Add the path to the database with the file's inode.
3. For each collated file in the database …
   a) For each of its paths in the database (matching inode), if the path does not exist in the file system, then delete the path from the database.
   b) If there are no paths left, delete the collated file from the database.
4. For each source path in the database, delete it from the database if it does not exist in the file system.
5. Finalise: close the database connection, write summary messages to the log.

4.10 List of Ruby methods

Method name and arguments	**Defined in**
CheckDir(dir, perm)	ruby_lib.rb
CheckFileSystemCollatedPathsAreInDB( )	clean.rb
CheckParameters( )	clean_omindex_log.rb
CheckParameters( )	ruby_lib.rb
Clean( )	clean_omindex_log.rb

CollateFiles( )	collate.rb
ConnectToDB( )	ruby_db_lib.rb
CopyFile(source_path, target_path)	ruby_lib.rb
CreateHardLink(original_path, new_path)	ruby_lib.rb
CreateTable(tableName, sql)	ruby_db_lib.rb
CreateTables	ruby_db_lib.rb
DeleteCollatedFileFromDB(inode)	ruby_db_lib.rb
DeleteCollatedFilePathsFromDbIfGone(inode)	ruby_db_lib.rb
DeleteCollatedPathFromDB(path)	ruby_db_lib.rb
DeleteSourceFileFromDB(path)	ruby_db_lib.rb
EnsureDirExists(dir)	ruby_lib.rb
ExistsTable?(tablename)	ruby_db_lib.rb
Finalise(exitcode, *msg)	clean.rb
Finalise(exitcode, *msg)	clean_omindex_log.rb
Finalise(exitcode, *msg)	collate.rb
Finalise(exitcode, *msg)	parse_cfg_for_bash.rb
GetAnyFileTypeExclusionReason(path)	clean.rb
GetChecksums(path)	ruby_lib.rb
GetCollationPathsFromTree(dir)	clean.rb
GetConfigFileData(fd)	ruby_lib.rb
GetInodeAndMtime(path)	ruby_lib.rb
GetMIMEtype(path)	ruby_lib.rb
Initialise	clean.rb
Initialise	clean_omindex_log.rb
Initialise	collate.rb
Initialise	parse_cfg_for_bash.rb
InitialiseParameters	ruby_lib.rb
InsertCollatedFileIntoDB(inode, md5, mime_type, mtime, sha1)	ruby_db_lib.rb

InsertCollatedPathIntoDB(inode, path)	ruby_db_lib.rb
InsertSourceFileIntoDB(inode, md5, mime_type, mtime, path, sha1)	ruby_db_lib.rb
IsCollatedPathInDB(path)	ruby_db_lib.rb
LogParameters( )	clean_omindex_log.rb
LogParameters( )	ruby_lib.rb
LookupAllCollatedFileInodes( )	ruby_db_lib.rb
LookupCollatedFileByChecksums(md5, sha1)	ruby_db_lib.rb
LookupCollatedFileByInode(inode)	ruby_db_lib.rb
LookupCollatedPathByInode(inode)	ruby_db_lib.rb
LookupCollatedPathsByInode(inode)	ruby_db_lib.rb
LookupSourceFileByPath(path)	ruby_db_lib.rb
LookupSourcePaths( )	ruby_db_lib.rb
NormaliseDir(dir)	ruby_lib.rb
NormaliseParameters( )	clean_omindex_log.rb
NormaliseParameters( )	ruby_lib.rb
NormalisePath(path)	ruby_lib.rb
ParseCommandLine( )	clean.rb
ParseCommandLine( )	clean_omindex_log.rb
ParseCommandLine( )	collate.rb
ParseCommandLine( )	parse_cfg_for_bash.rb
ParseConfigFile(config_path, *valid_keywords)	ruby_lib.rb
ProcessCollatedPath(path)	clean.rb
ProcessFile(source_path)	collate.rb
ShellEscape(str)	ruby_lib.rb
StrToTime(str)	ruby_lib.rb
UpdateCollatedFileInDB(inode, md5, mime_type, mtime, sha1)	ruby_db_lib.rb
UpdateSourceFileInDB(inode, md5, mime_type, mtime, path, sha1)	ruby_db_lib.rb

Usage(verbosity)	clean.rb
Usage(verbosity)	clean_omindex_log.rb
Usage(verbosity)	collate.rb
Usage(verbosity)	parse_cfg_for_bash.rb
WriteParametersForBash( )	parse_cfg_for_bash.rb

Note: the raw data for the table was generated by:

```
grep '^def' *.rb | awk -F: '{print $2 "\t" $1}' | sed 's/^def //' | sort
```

4.11 MIME types

File MIME types are determined and recorded in the database but not used. This anticipates Xapian Omega's omindex utility starting to use MIME types instead of file name extensions for some file types. This change was being discussed in the Xapian mailing list at the time docoll 0.7.2 was being developed.

5 Configuration files

Configuration files are in the /etc/opt/docoll/<instance> directories, for example:

```
docoll@CW8vDS:~$ ls -l /etc/opt/docoll/default/*.cfg
-rw-r--r-- 1 docoll docoll  263 Sep 24  2011 /etc/opt/docoll/default/backup_db.cfg
-rw-r--r-- 1 docoll docoll 1692 Mar 29 09:54 /etc/opt/docoll/default/clean_omindex_log.cfg
-rw-r--r-- 1 docoll docoll 3275 Mar 30 21:59 /etc/opt/docoll/default/collate.cfg
-rw-r--r-- 1 docoll docoll 5607 Mar 30 20:10 /etc/opt/docoll/default/omindex.sh.cfg
-rw-r--r-- 1 docoll docoll 1994 Mar 30 20:10 /etc/opt/docoll/default/run_scripts.cfg
```

The example listing above shows:

- **backup_db.cfg**, the configuration file to set backup retention time for script backup_db.sh.
- **clean_omindex_log.cfg**, the configuration file used by clean_omindex_log.rb.
- **collate.cfg**, the configuration file used by scripts backup_db.sh, clean.rb, collate.rb and omindex.sh.
- **omindex.sh.cfg**, used by script omindex.sh (it uses two configuration files).
- **run_scripts.cfg**, the configuration file used by run_scripts.sh.

6 Logs

run_scripts.sh' configuration file includes its output directory. It writes its log in the output directory's log subdirectory.

The other scripts have a mandatory command line option giving their log file name. omindex.sh also makes the Xapian utility omindex log to a separate file (because omindex output is huge) in the same directory as its own log.

When the scripts are run in the normal way by run_scripts.sh, run_scripts.sh runs each of them with a command line option to log to a file in the same directory as its own log.

Errors detected by bash or the Ruby interpreter are written to stderr. In the normal case (run_scripts.sh being run by cron and the bash error not in run_scripts.sh) they appear in the run_scripts.sh log.

7 Development, test etc. instances

Development, test etc. docoll instances may be set up just like any other instance. Details in related document "docoll server sysadmin guide".

8 Packaging

Packaging is described in related document "docoll packagers guide".

9 Appendix – 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.