Xapian developer guide1.4

Release 1.4.3

Xapian Documentation Team
Contributors

March 14, 2019

Contents

Xapian is an open source search engine library, which allows developers to add advanced indexing and search facilities to their own applications. This manual aims to explain how to work on and contribute to Xapian itself; if you want to use Xapian in your own project, you should look at our Xapian user manual.

Note: This is very early days for this guide, so please let us know any issues you spot or how we can improve it in any way. There's a lot of additional information about developing for Xapian in the HACKING file in xapian-core which we hope to move here in future, and in the meantime that's a good place to look for information on writing and running tests, debugging, and options available when configuring the source tree. It also contains our documentation for making a Xapian release.

Since Xapian is an open source project, it is entirely dependent on contributions – many of them from people like you! These contributions come in many forms, from spotting when some documentation isn't clear and maybe improving it, through fixing bugs up to designing and implementing completely new features. We need all types of contribution for Xapian to continue to evolve and serve its users.

This guide is intended to:

- help you understand how to make and contribute changes to Xapian
- lay out some of our "rules of the road", which are there to make it easier for everyone to collaborate on Xapian

We recommend you at least skim through all of this guide, so you know what information is available to you. However, if you absolutely must jump in head first, you should at least read our *advice for new contributors*.

Todo

Talk about how to manage security issues.

CHAPTER 1

Getting in touch

The Xapian community typically works "in the open", via mailing lists and an IRC channel:

- Our mailing lists are open for anyone to join, although (because we don't want to relay spam to everyone) if you aren't subscribed to the list someone will have to manually approve your message. Please be patient and don't resend a message just because it doesn't appear right away.
- We're on #xapian on `irc.freenode.net`. IRC is a simple text chat system where many members of the community hang out; although because we're distributed around the world, you may not get an instant response. That doesn't mean we're ignoring you, so either hang on for a reply, or you can use the mailing list instead.

CHAPTER 2

Contents

2.1 Getting started

Note: Currently this guide is written assuming you are either developing on something like Unix (probably either Linux or OS X). You can use software suchas Virtual Box to run a 'virtual' Linux machine on another operating system; in this case we recommend using the most recent Ubuntu LTS release.

It should be possible to build and develop Xapian on Windows, but we currently don't have any documentation on doing so, or any active developers with suitable experience.

2.1.1 Getting the source code

First off, let's make sure you have a copy of the Xapian source and can build it and run the tests. This is generally a little different to if you're just installing Xapian to use it, because you'll be working with the entire source tree rather than individual pieces. The Xapian build system has some support for this, but let's get you a copy of everything first:

```
$ git clone git://git.xapian.org/xapian
$ cd xapian
```

This will 'clone' a complete copy of the Xapian source code, including not only the core library but also the variable language bindings (for use from Python, Lua, Ruby and so on) and the self-contained web search system 'Omega'. It also contains all the tests for those various components.

2.1.2 Installing the dependencies

Debian / Ubuntu

For a recent version of Debian or Ubuntu, this command should ensure you have all the necessary tools and libraries:

```
$ apt-get install build-essential m4 perl python zlib1g-dev uuid-dev \
  wget bison tcl libpcre3-dev libmagic-dev valgrind ccache eatmydata \
  doxygen graphviz help2man python-docutils pngcrush python-sphinx \
  python3-sphinx mono-devel openjdk-8-jdk lua5.2 liblua5.2-dev \
  php-dev php-cli python-dev python3-dev ruby-dev tcl-dev texinfo
```

OS X

You need to install Apple's XCode tools, which contain their compiler, debugger and various other tools. You can do that from within the AppStore.

We recommend using homebrew to install and manage additional libraries and tools on OS X. Once you've installed XCode and homebrew, you can get all the dependencies you need for Xapian using:

```
$ brew install libmagic pcre \
  lua mono perl php python python3 ruby tcl-tk \
  doxygen help2man graphviz pngcrush
# and some python-specific documentation tools
$ pip install sphinx docutils
$ pip3 install sphinx
```

(We install documentation tools for both python2 and python3, in the same way we build the bindings for both of them.)

2.1.3 Building Xapian

Bootstrapping the code

Xapian needs to set up a few things with a fresh clone of the code, as well as downloading and building some tools for which we require very precise versions. You should run this command in the `xapian` directory that was created earlier when you cloned the source code:

```
$ ./bootstrap
```

To download tools, bootstrap will use `wget`, `curl` or `lwp-request` if installed. If not, it will give an error telling you the URL to download from by hand and where to copy the file to.

Note: As well as installing some tools, bootstrap will also run `autoreconf` on each of the checked-out subdirectories, and generate a top-level `configure` script. This configure script allows you to configure xapian-core and any other modules you've checked out with a single simple command, such that the other modules link against the uninstalled xapian-core (which is very handy for development work and a bit fiddly to set up by hand). It automatically passes `--enable-maintainer-mode` to the subprojects so that the autotools will be rerun if `configure.ac`, `Makefile.am`, etc are modified.

Warning: If you are tracking development in git, there will sometimes be changes to the build system sources which require regeneration of the generated makefiles and associated machinery. We aim to make the build system automatically regenerate the necessary files, but in the event that a build fails after an update, it may be worth re-running the bootstrap script to regenerate the build system from scratch, before looking for the cause of the error elsewhere.

Configuring the code

Configuring the code is mostly about Xapian's build system automatically detecting where all its dependencies are on your computer, so it knows how to use them. However there are various options that allow you to either override the autodetection (for instance if you wanted to build python bindings against a particular version of python) or change some defaults. For now, however, we'll just run it accepting all its defaults:

```
$ ./configure
```

Note that on OS X you probably want to turn off the Perl and TCL8 bindings when developing, as there are some complexities when developing against the system versions, and the homebrew versions are slightly awkward:

```
$ ./configure --without-perl --without-tcl
```

Building Xapian

Building Xapian is just a matter of typing:

```
$ make
```

First it will build xapian-core, the core library. Then it will build Omega and the language bindings, using the version of xapian-core you've just built, but not yet installed. (This is the bit that causes some problems on OS X if you use system versions of any of the languages.)

2.1.4 Running the tests

Xapian has a comprehensive test suite, and it's a good idea to get into the habit of running it. From the top of the clone, just run:

```
$ make check
```

Again, the tests for xapian-core are run first, then Omega and then the language bindings. If any test fails, the build system will stop there.

2.1.5 Summary

Now you've got everything working, you probably want to look at writing code, or if you're trying to fix a bug then you might want to learn about debugging Xapian.

Todo

The other sections of the manual haven't been written yet, so this part isn't terribly helpful. Sorry!

2.2 Coding and other conventions in Xapian

We aim for Xapian to be:

- *Cross-platform*: this means it will compile and run on a range of different platforms. We have three sets of automated build systems to help us keep track of this: the Xapian buildbots, our builds on Travis CI, and finally builds on AppVeyor. The last two will be triggered automatically if you submit changes using *a pull request on github*.
- *Able to build "cleanly"*, meaning without warnings, on a range of compilers. Note that the two main C++ compilers in use these days are from clang and GCC, and they have slightly different sets of warnings and behaviours. Our automated builds run against both compilers. When you build Xapian, the compiler configuration used by default will highlight warnings and refuse to complete the build if it finds any.

- *Documented* and *tested* throughout. Although not all of it is fully-documented or tested as it stands, if we add documentation and tests every time we add a feature, fix a bug, or work on existing code, then we will keep on improving.

  If you'd like to improve our test code coverage, our code coverage report may be helpful in choosing something to add tests for. Remember that code coverage isn't the same as having useful tests for the code, so even in parts of the codebase that have good coverage there may be new tests worth writing.

With that in mind, we have some conventions and standards that we try to adhere to. It's difficult to get away from big lists of things to do and not do, but we've tried to explain why as much as possible, and to group things in a way that makes it easier to find useful information. It's a good idea to at least skim through this material so you can go back for a detailed look when you need.

2.2.1 C++ conventions

Code layout

Indentation

Indent C++ code by 4 spaces for a new indentation level, and set your editor to tab-fill indentation (with a tab being 8 spaces wide).

As an exception, "public", "protected" and "private" declarations in classes and structs should be indented by 2 spaces, and the following code should be indented by 2 more spaces:

```
class Foo {
  public:
    method();
};
```

The rationale for this exception is that class definitions in header files often have fairly long lines, so losing an indent level to the access specifier tends to make class definitions less readable.

The default access for a class is always "private", so there's no need to specify that explicitly - in other words, write this:

```
class Foo {
    int internal_method();

  public:
    int external_method();
};
```

Don't write this:

```
class Foo {
  private:
    int internal_method();

  public:
    int external_method();
};
```

If a class only contains public methods and data, consider declaring it as a "struct" (the only difference in C++ is that the default access for a struct is "public").

Spaces and line breaks

Put a space before the `(` after control flow constructs like `for`, `if`, `while`, and so on. So write `if (strlen(p) > 10)` not `if(strlen (p) > 10)`. Don't put a space before the `(` in function calls.

When `if`, `else`, `for`, `while`, `do`, `switch`, `case`, `default`, `try`, or `catch` is followed by a block enclosed in braces, the opening brace should be on the same line, like so:

```
if (x > 12) {
    foo(x);
    x = 12;
} else {
    bar(x);
}
```

The rationale for this is that it conserves vertical space (allowing more code to fit on screen) without reducing readability.

C++ idioms in Xapian

- If you have an empty loop body, use `{ }` rather than `;` as the former stands out more clearly to the reader (but also consider if the code might be clearer written a different way).
- Prefer `++i;` to `i++;`, `i += 1;`, or `i = i + 1`. For simple integer variables these should generate equivalent (if not identical) code, but if i is an iterator object then the pre-increment form can be more efficient in some cases with some compilers. It's simpler and more consistent to always use the pre-increment form (unless you make use of the old value which the post-increment form returns). For the same reasons, prefer `--i;` to `i--;`, `i -= 1;`, or `i = i - 1;`.
- Prefer `container.empty()` to `container.size() == 0` (and `!container.empty()` to `container.size() != 0` or `container.size() > 0`).

  Some containers (e.g. `std::forward_list`) support `empty()` but not `size()`. Pre-C++11 finding the size of a container wasn't necessarily a constant time operation for some containers (e.g. `std::list` with GCC) - that's no longer the case for any STL containers since C++11, but it could still be true for non-STL containers.

  Also the `empty()` form makes the intent of the test more explicit.
- Prefer not to use `else` when the control flow is diverted elsewhere at the end of the `if` block (e.g. by `return`, `continue`, `break`, `throw`). This eliminates a level of indentation from the code in the `else` block, and typically makes the control flow logic clearer. For example:

```
    if (x == 0) {
        foo();
        return;
    }

    while (x--) {
        bar();
    }
```

  rather than:

```
    if (x == 0) {
        foo();
        return;
    } else {
        while (x--) {
```

```
            bar();
        }
    }
```

- For standard ISO C headers, prefer the C++ form for ISO C headers (e.g. `#include <cstdlib>` rather than `#include <stdlib.h>`) unless there's a good reason (e.g. portability) to do otherwise. Be sure to document such exceptions to avoid another developer changing them to the standard form. Global exceptions: <signal.h> (lots of POSIX stuff which e.g. Sun's compiler doesn't provide in <csignal>).
- For standard ISO C++ headers, *always* use the ISO C++ form `#include <list>` (pre-ISO compilers used `#include <list.h>`, but GCC has generated a warning for this form for years, and GCC 4.3 dropped support entirely).
- Prefer `new SomeClass` to `new SomeClass()`, since the latter tends to lead one to write `SomeClass foo();` which is a function prototype, and not equivalent to the variable definition `SomeClass foo`. However, note that `new SomePODType()` is *not* the same as `new SomePODType` (if SomePODType is a Plain Old Data type) - the former will zero-initialise scalar members of SomePODType.
- RTTI (`dynamic_cast<>`, `typeid`, `std::typeinfo`): Needing to use RTTI features in the library most likely indicates a design flaw, and you should avoid use of these features. Where necessary, you can use a technique similar to `Database::as_networkdatabase()` to replace `dynamic_cast<>`.
- `using namespace std;` and `using std::XXX;` - it's OK to use these in applications, library code, and internal library headers. But in externally visible headers (such as anything included by `#include <xapian.h>`) you *MUST* use explicit `std::` qualifiers - it's not acceptable to pull anything from namespace std into the namespace of an application which uses Xapian.
- Use C++ style casts (`static_cast<>`, `reinterpret_cast<>`, and `const_cast<>`) or constructor-syntax (e.g. `double(value)`) in preference to C style casts. The syntax of the C++ casts is ugly, but they do make the intent much clearer which is definitely a good thing, and they avoid issues such as casting away const when you only meant to cast the type of a pointer.
- `std::pair<>` with an STL class as one (or both) of the members can produce very long symbols (over 4KB!) after name mangling - long enough to overflow the size limits of some vendor compilers or toolchains (so this can affect GCC if it is using the system ld or as). Even where the compiler works, the symbol bloat in an unstripped build is probably best avoided, so it's preferable to use a simple two member struct instead. The code is probably more readable anyway, and easier to extend if more members are needed later.
- We try to avoid putting the full definition of virtual methods in header files. This is because current compilers can't (as far as we know) inline virtual methods, so putting the definition in the header file simply slows down compilation (and, because method definitions often require further header files to be included, this can result in many more files needing recompilation after a change to a header file than is really necessary). Just put the declaration in the header file, and put the definition in a .cc file with the same basename.

Efficient use of `std::string`

- When passing an empty string to a method expecting `const std::string &` prefer `std::string()` to `""` or `std::string("")` (it is more efficient with some compilers).
- To make a string object empty, `s.resize(0)` (if you want to keep the current reserved space) or `s = string()` (if you don't) seem the best options.
- Use `std::string::assign()` rather than building a temporary string object and assigning that. For example, `foo = std::string(ptr, len);` is better written as `foo.assign(ptr, len);`.
- It's generally better to build up strings using += rather than combining series of components with +. So `foo = a + " and " + c` is better written as `foo = a; foo += " and "; foo += c;`. It's possible for compilers to handle the former without a lot of temporary string objects by returning a proxy object to allow the

concatenation to happen lazily, but not all compilers do this, and it's likely to still have some overhead. Note that GCC 4.1 seems to produce larger code in some cases for the latter approach, but it's a definite win with GCC 4.4.

- `std::string(1, '\0')` seems to be slightly more efficient than `std::string("", 1)` for constructing a `std::string` containing a single ASCII nul character.

Use of C++ Features

C++11

As of Xapian 1.3.3, a compiler with decent support for C++11 is required to build Xapian. We currently aim to allow users to use a non-C++11 compiler to build code which uses Xapian.

There are now several compilers with good C++11 support, but there are a few shortfalls in commonly deployed versions of most of them. Often we can work around this, and we should do where the effort is low compared to the gain (so a compiler version which is widely used is more worth supporting than one which is hardly used by anyone).

However, we shouldn't have to jump through hoops to cater for compilers where their authors aren't putting in the effort to keep up with the language standards.

Please avoid the following C++11 features for the time being:

- `std::to_string()` - this is completely missing on current versions of mingw and cygwin - in the library, you can `#include "str.h"` and then use the `str()` function instead for most cases. This is also usually faster than `std::to_string()`.

C++ features we assume

We assume that all compilers will correctly implement the following, so it's safe to rely on them when working on Xapian.

- We assume `<sstream>` is available. GCC < 2.95.3 didn't have it but GCC 2.95.3 includes a backported version. We aren't aware of any other compilers still in use which lack it.
- Non-".h" versions of standard ISO C++ headers (e.g. `#include <list>` rather than `#include <list.h>`). We aren't aware of any compiler still in use which lacks these, and GCC 4.3 no longer has the old versions. If there are any, we could add a directory full of forwarding headers to work around this issue.
- Standard header `<limits>` (for `numeric_limits<>`) - for GCC, this was added in GCC 3.0.
- Standard header `<streambuf>` (GCC < 3.0 only has `<streambuf.h>`).

Exceptions

When catching an exception which is an object, do it by const reference, so like this:

```
try {
    foo();
} catch (const ErrorClass& e) {
    bar(e);
}
```

Catching by value is bad because it "slices" the object if an object of a derived type is thrown. Even if derived types aren't a worry, it also causes the copy constructor to be called needlessly. More information is available in a Standard C++ FAQ entry.

A const reference is preferable to a non-const reference as it stops the object being inadvertently modified. In the rare cases when you want to modify the caught object, a non-const reference is OK.

Exceptions should be avoided except for truly exceptional situations, since throwing and handling them has a significant cost. It also generally makes the API easier to understand, and client code easier to read.

Include ordering for source files

To help us move towards a consistent ordering of `#include` lines in source files, please follow the following policy when ordering them:

- `#include <config.h>` should be first, and use <> not "" (as recommended by the autoconf manual). Always include `config.h` from C/C++ source files, but don't include it from header files - the autoconf manual recommends that it should be included first, so including it from headers is either redundant, or may hide a missing config.h include in the source file the header was included from (better to get an error in this case).
- The header corresponding to the source file should be next. This means that compilation of the library ensures that each header with a corresponding source file is "self supporting" (i.e. it implicitly or explicitly includes all of the headers it requires).
- External xapian-core headers, alphabetically. When included from other external headers, use <> to reduce problems with finding headers in the user's source tree by mistake. In sources and internal headers, use "" (?) - practically this makes no difference as we have `-I` for srcdir and builddir, but <> suggests installed header files so "" seems more natural).
- Internal headers, alphabetically (using "").
- "Safe" versions of library headers (include these first to avoid issues if other library headers include the ones we want to wrap). Use "" and order alphabetically.
- Library headers, alphabetically.
- Standard C++ headers, alphabetically. Use the modern (no .h suffix) names.

Branch Prediction Hints

For compilers which support `__builtin_expect()` (GCC >= 3.0 and some others) you can provide manual hints to assist branch prediction. We've wrapped these in macros which evaluate to just their argument for compilers which don't support `__builtin_expect()`.

Within the xapian-core library code, you can mark the expressions in `if` and `while` statements as `rare` (if the condition is rarely true) or `usual` (if the condition is usually true).

For example:

```
if (rare(something_unusual())) deal_with_it();

while (usual(!end_condition()) keep_going();
```

It's easy to make incorrect assumptions about where hotspots are and which branches are usually taken or not, so except for really obvious cases (such as `if (!consistency_check()) throw_exception();`) you should benchmark that new `rare` and `usual` hints help rather than hinder before committing them to the repository. It's also likely to be a waste of effort to add them outside of areas of code which are executed very frequently.

Don't expect miracles - the first 15 uses added saved approximately 1%.

If you know how to implement the `rare` and `usual` macros for other compilers, please let us know.

Use of Assert

Use Assert to perform internal consistency checks, and to check for invalid arguments to functions and methods (e.g. passing a NULL pointer when this isn't permitted). It should *NOT* be used to check for error conditions such as file read errors, memory allocation failing, etc (since we want to perform such checks in non-debug builds too).

File format errors should also not be tested with Assert - we want to catch a corrupted database or a malformed input file in a non-debug build too.

There are several variants of Assert:

- `Assert(P)` – asserts that expression P is true.
- `AssertRel(a,rel,b)` – asserts that (a rel b) is true - rel can be a boolean relational operator, i.e. one of `==`, `!=`, `>`, `>=`, `<`, `<=`. The message given if the assertion fails reports the values of a and b, so `AssertRel(a,<,b);` is more helpful than `Assert(a < b);`
- `AssertEq(a,b)` – shorthand for `AssertRel(a,==,b)`.
- `AssertEqDouble(a,b)` – asserts a and b differ by less than `DBL_EPSILON`
- `AssertParanoid(P)` – a particularly expensive assertion. If you want a build with Asserts enabled, but without a great performance overhead, then passing –enable-assertions=partial to configure and AssertParanoids won't be checked, but Asserts will. You can also use `AssertRelParanoid` and `AssertEqParanoid`.

An earlier assert, `CompileTimeAssert(P)`, has now been removed, since we require C++11 support from the compiler, and C++11 added `static_assert`.

2.2.2 Portability

C++ Portability Issues

Web Resources

The C++ Super-FAQ covers many frequently asked C++ questions.

Header Portability Issues

`<fcntl.h>`

> Don't directly `#include <fcntl.h>` - instead `#include "safefcntl.h"`.
>
> The main reason for this is that when using certain compilers on certain versions of Solaris, fcntl.h does `#define open open64`. Sadly this breaks C++ code which has methods called open (as we do). There's a cunning workaround for this problem in `common/safefcntl.h`.
>
> Also, `safefcntl.h` ensures the `O_BINARY` is defined (to 0 if not required) so calls to `open()` and `creat()` can specify `O_BINARY` unconditionally for the benefit of platforms which discriminate between text and binary files.

`<windows.h>`

> Don't directly `#include <windows.h>` - instead `#include "safewindows.h"` which reduces the bloat of header files included and prevents some of the more egregious namespace pollution. It also defines any constants we need which might be missing in older versions of the mingw headers.

`<winsock2.h>`

Don't directly `#include <winsock2.h>` - instead `#include "safewinsock2.h"`. This ensure that safewindows.h is included before <winsock2.h> to avoid winsock2.h including windows.h without our namespace pollution reducing workarounds.

`<sys/select.h>`

Don't directly `#include <sys/select.h>` - instead `#include "safesysselect.h"` which supports older UNIX platforms which predate POSIX 1003.1-2001 and works around a problem on Solaris.

`<sys/socket.h>`

Don't directly `#include <sys/socket.h>` - instead `#include "safesyssocket.h"` which supports older UNIX platforms which predate POSIX 1003.1-2001 and works on Windows too.

`<sys/stat.h>`

Don't directly `#include <sys/stat.h>` - instead `#include "safesysstat.h"` which under MSVC enables stat to work on files > 2GB, defines the missing POSIX macros S_ISDIR and S_ISREG, pulls in <direct.h> for mkdir() (which is provided by sys/stat.h under UNIX) and provides a compatibility wrapper for mkdir() which takes 2 arguments (so code using mkdir can always just pass two arguments).

`<sys/wait.h>`

To get *WEXITSTATUS* or *WIFEXITED* defined, `#include "safesyswait.h"`. Note that this won't provide *waitpid()*, etc on Microsoft Windows, since these functions are only really useful to use when *fork()* is available.

`<unistd.h>`

Don't directly `#include <unistd.h>` - instead `#include "safeunistd.h"` – MSVC doesn't even HAVE unistd.h!

The various "safe" headers are maintained in `xapian-core/common`, but also used by Omega. Currently bootstrap sorts out setting up a copy of this subdirectory via a secondary git checkout.

Warning-Free Compilation

Compiling without warnings on every platform is our goal, though it's not always possible to achieve. For example, some GCC 3.x compilers produce the occasional bogus warning (e.g. warning that a variable may be used uninitialised, despite it being initialised at the point of declaration!)

You should consider configure-ing with:

```
./configure CXXFLAGS=-Werror
```

when doing development work on Xapian. This promotes warnings to errors, which should ensure you at least don't introduce new warnings for the compiler you're using.

If you configure with `--enable-maintainer-mode`, and are using GCC 4.1 or newer, this is done for you automatically. This is intended to be an aid rather than a form of automated punishment - it's all too easy to miss a new warning as once a file is compiled, you don't see it unless you modify that file or one of its dependencies.

With Intel's C++ compiler, `--enable-maintainer-mode` also enables `-Werror`. If you know the equivalent of `-Werror` for other compilers, please add a note here, or tell us so that we can add a note.

Miscellaneous Portability Issues

Make sure that the last line of any source file ends with a linefeed character since it's undefined behaviour if it doesn't (most compilers accept it, though at least GCC gives a warning).

Makefile Portability

We don't want to force those building Xapian from the source distribution to have to use GNU make. Requiring GNU make for "make dist" isn't such a problem but it's probably better to use portable constructs everywhere to avoid problems when people move or copy code between targets. If you do make use of non-portable constructs where it's OK, add a comment noting the special circumstances which justify doing so.

Here's an incomplete list of things to avoid:

- Don't use "$(RM)" - it's defined by GNU make, but using it actually harms portability as other makes don't define it. Use plain "rm" instead.
- Don't use "%" pattern rules - these are GNU make specific. Use an implicit rule (e.g. ".c.o:") if you can. Otherwise, write out each version explicitly.
- Don't use "$<" except in implicit rules. This is an annoying restriction, as using "$<" makes it much easier to make VPATH builds work. But it's only portable in implicit rules. Tips for rewriting - if it's a source file, write it as:

```
$(srcdir)/foo.ext
```

  If it's a generated object file or similar, just write the name as is. The tricky case is a generated file which isn't in git but is shipped in the distribution tarball, as such a file could be in either the source or build tree. Use this trick to make sure it's found whichever directory it's in:

```
`test -f foo.ext || echo '$(srcdir)/'`foo.ext
```

- Don't use "exit 0" to make a rule fail. Use "false" instead. BSD make doesn't like "exit 0" in a rule.
- Don't use make conditionals. Automake offers conditionals which may be of use, and these are implemented to work with any make. See the automake manual for details, and a few caveats.
- The list of portable utilities is:

```
cat cmp cp diff echo egrep expr false grep install-info
ln ls mkdir mv pwd rm rmdir sed sleep sort tar test touch true
```

  Note that versions of these (GNU versions in particular) support switches which aren't portable - notably, "test -r" isn't portable; neither is "cp -a". And note that "mkdir -p" isn't portable - the semantics vary. The autoconf manual has some useful information about writing portable shell code (most of it not specific to autoconf):

```
https://www.gnu.org/software/autoconf/manual/autoconf.html#Portable-Shell
```

- Don't use "include" - it's not present in BSD make (at least some versions have ".include" instead, but that doesn't really seem to help...) Automake provides a configure-time include, which may provide a replacement for some uses of "include".
- It appears that BSD make only supports VPATH for implicit rules (e.g. ".c.o:") - there's certainly a restriction there which is not present in GNU make. We used to try to work around this, but now we use AM_MAINTAINER_MODE to disable rules which are only needed by those developing Xapian (these were the rules which caused problems). And we recommend those developing Xapian use GNU make to avoid problems.
- Rules with multiple targets can cause problems for parallel builds. These rules are really just a shorthand for multiple rules with the same prerequisites and commands, and it is fine to use them in this way. However, a common temptation is to use them when a single invocation of a command generates multiple output files, by adding each of the output files as a target. Eg, if a swig language module generates xapian_wrap.cc and xapian_wrap.h, it is tempting to add a single rule something like:

```
# This rule has a problem
xapian_wrap.cc xapian_wrap.h: xapian.i
        SWIG_commands
```

This can result in SWIG_commands being run twice, in parallel. If SWIG_commands generates any temporary files, the two invocations can interfere causing one of them to fail.

Instead of this rule, one solution is to pick one of the output files as a primary target, and add a dependency for the second output file on the first output file:

```
# This rule also has a problem
xapian_wrap.h: xapian_wrap.cc
xapian_wrap.cc: xapian.i
        SWIG_commands
```

This ensures that make knows that only one invocation of SWIG_commands is necessary, but could result in problems if the invocation of SWIG_commands failed after creating xapian_wrap.cc, but before creating xapian_wrap.h. Instead, we recommend creating an intermediate target:

```
# This rule works in most cases
xapian_wrap.cc xapian_wrap.h: xapian_wrap.stamp
xapian_wrap.stamp: xapian.i
        SWIG_commands
        touch $@
```

Because the intermediate target is only touched after the commands have executed successfully, subsequent builds will always retry the commands if an error occurs. Note that the intermediate target cannot be a "phony" target because this would result in the commands being re-run for every build.

However, this rule still has a problem - if the xapian_wrap.cc and xapian_wrap.h files are removed, but the xapian_wrap.stamp file is not, the .cc and .h files will not be regenerated. There is no simple solution to this, but the following is a recipe taken from the automake manual which works. For details of *why* it works, see the section in the automake manual titled "Multiple Outputs":

```
# This rule works even if some of the output files were removed
xapian_wrap.cc xapian_wrap.h: xapian_wrap.stamp
## Recover from the removal of $@.  A full explanation of these rules is in
## the automake manual under the heading "Multiple Outputs".
        @if test -f $@; then :; else \
          trap 'rm -rf xapian_wrap.lock xapian_wrap.stamp' 1 2 13 15; \
          if mkdir xapian_wrap.lock 2>/dev/null; then \
            rm -f xapian_wrap.stamp; \
            $(MAKE) $(AM_MAKEFLAGS) xapian_wrap.stamp; \
            rmdir xapian_wrap.lock; \
          else \
            while test -d xapian_wrap.lock; do sleep 1; done; \
            test -f xapian_wrap.stamp; exit $$?; \
          fi; \
        fi
xapian_wrap.stamp: xapian.i
        SWIG_commands
        touch $@
```

- This is actually a robustness point, not portability per se. Rules which generate files should be careful not to leave a partial file in place if there's an error as it will have a timestamp which leads make to believe it's up-to-date. So this is bad:

```
foo.cc: script.pl
        $PERL script.pl > foo.cc
```

This is better:

```
foo.cc: script.pl
        $PERL script.pl > foo.tmp
        mv foo.tmp foo.cc
```

Alternatively, pass the output filename to the script and make sure you delete the output on error or a signal (although this approach can leave a partial file in place if the power fails). All used Makefile.am-s and scripts have been checked (and fixed if required) as of 2003-07-10 (didn't check xapian-bindings).

- Another robustness point - if you add a non-file target to a makefile, you should also list it in ".PHONY". Otherwise your target won't get remade reliably if someone creates a file with the same name in their tree. For example:

```
.PHONY: hello goodbye

hello:
      echo hello

goodbye:
      echo goodbye
```

And lastly a style point - using "@" to suppress echoing of commands being executed removes choice from the user - they may want to see what commands are being executed. And if they don't want to, many versions of make support the use "make -s" to suppress the echoing of commands.

Using `@echo` on a message sent to stdout or stderr is acceptable (since it avoids showing the message twice). Otherwise don't use "@" - it makes it harder to track down problems in the makefiles.

2.2.3 Other conventions

Configure Options

Especially for a library, compile-time options aren't a good solution for how to integrate a new feature. An increasingly large number of users install pre-built binary packages rather than building from source, and unless the package is capable of being split into modules, the packager has to choose a set of compile-time options to use. And they'll tend to choose either the standard ones, or perhaps a broader set to try to keep everyone happy. For a library, similar issues occur when installing from source as well - the sysadmin must choose the options which will keep all users happy.

Another problem with compile-time options is that it's hard to ensure that a change doesn't break compilation under some combination of options without actually building and running the test-suite on all combinations. The fewer compile-time options, the more likely the code will compile with every combination of them.

So please think carefully before adding more compile-time options. They're probably OK for experimental features (but should go away once a feature is no longer experimental). Options to instrument a build for special purposes (debug, profiling, etc) are also acceptable. Disabling whole features probably isn't (e.g. the `--disable-backend-XXX` options we already have are dubious, though being able to disable the remote backend can be useful when trying to get Xapian going on a platform).

Naming of Scripts

Scripts generally should *not* have an extension indicating the language they are currently implemented in (e.g. `runtest` rather than `runtest.sh` or `runtest.pl`). The problem with such an extension is that if we decide to reimplement the script in a different language, we either have to rename the script (which is annoying as people will be used to the name, and may have embedded it in their own scripts), or we have a script with a confusing name (e.g. a Python script with extension `.pl`).

The above reasoning doesn't apply to scripts which have to be in a particular language for some reason, though for consistency they probably shouldn't get an extension either, unless there's a good reason to have one.

2.2.4 Marking Features as Deprecated

In the API headers, a feature (a class, method, function, enum, typedef, etc) can be marked as deprecated by using the `XAPIAN_DEPRECATED()` or `XAPIAN_DEPRECATED_CLASS` macros. Note that you can't deprecate a preprocessor macro.

For compilers with a suitable mechanism (such as GCC, clang and MSVC) this causes compile-time warning messages to be emitted for any use of the deprecated feature. For compilers without support, the macro just expands to its argument.

Sometimes a deprecated feature will also be removed from the library itself (particularly something like a typedef), but if the feature is still used inside the library (for example, so we can define class methods), then use `XAPIAN_DEPRECATED_EX()` or `XAPIAN_DEPRECATED_CLASS_EX` instead, which will only issue a warning in user code (this relies on user code including xapian.h and library code including individual headers)

You must add this line to any API header which uses `XAPIAN_DEPRECATED()` or `XAPIAN_DEPRECATED_CLASS`:

```
#include <xapian/deprecated.h>
```

When marking a feature as deprecated, document the deprecation in `docs/deprecation.rst`. When actually removing deprecated features, please tidy up by removing the inclusion of `<xapian/deprecated.h>` from any file which no longer marks any features as deprecated.

The `XAPIAN_DEPRECATED()` macro should wrap the whole declaration except for the semicolon and any "definition" part, for example:

```
XAPIAN_DEPRECATED(int old_function(double arg));

class Foo {
  public:
    XAPIAN_DEPRECATED(int old_method());

    XAPIAN_DEPRECATED(int old_const_method() const);

    XAPIAN_DEPRECATED(virtual int old_virt_method()) = 0;

    XAPIAN_DEPRECATED(static int old_static_method());

    XAPIAN_DEPRECATED(static const int OLD_CONSTANT) = 42;
};
```

Mark a class as deprecated by inserting `XAPIAN_DEPRECATED_CLASS` after the class keyword like so:

```
class XAPIAN_DEPRECATED_CLASS Foo {
  public:
    Foo() { }

    // ...
};
```

You can simply mark a method defined inline in a class with `XAPIAN_DEPRECATED()` like so:

```
class Foo {
  public:
    // This failed to compile with GCC 3.3.5.
```

```
    XAPIAN_DEPRECATED(int old_inline_method()) { return 42; }
};
```

2.2.5 API Structure Notes

We use reference counted pointers for most API classes. These are implemented using `Xapian::Internal::intrusive_ptr`, the implementation of which is exposed for efficiency, and because it's unlikely we'll need to change it frequently, if at all.

For the reference counted classes, the API class (e.g. `Xapian::Enquire`) is really just a wrapper around a reference counted pointer. This points to an internal class (e.g. `Xapian::Enquire::Internal`). The reference counted pointer is a member variable of the API class called internal. Conceptually this member is private, though it typically isn't declared as private (this is to avoid littering the external headers with friend declarations for non-API classes).

There are a few exceptions to the reference counted structure, such as `MSetIterator` and `ESetIterator` which have an exposed implementation. Tests show this makes a substantial difference to speed (it's ~20% faster) in typical cases of iterator use.

The postfix `operator++` for iterators should be implemented inline in terms of the prefix form as described by Joe Buck on the gcc mailing list:

```
    class some_iterator {
    public:
        // ...
        some_iterator& operator++();
        some_iterator operator++(int) {
            some_iterator tmp = *this;
            operator++();
            return tmp;
        }
    };
```

> The compiler is allowed to assume that the copy constructor only does a copy, and to optimize away unneeded copy operations. The result in this case should be that, for some_iterator above, using the postfix operator without using the result should give code equivalent to using the prefix operator.
>
> [With modern compilers], you should find that this style comes very close to eliminating any penalty from "incorrect" use of the postfix form.

Xapian's PostingIterator, TermIterator, PositionIterator, and ValueIterator all have only one data member which fits in a register.

2.3 Contributing to Xapian

Xapian is an open source project, depending on a community of volunteers. There are lots of ways of contributing to Xapian:

- Join our mailing lists or IRC channel (#xapian on Freenode) and help people out.
- Talk or write about Xapian. By explaining how you've used Xapian you not only help spread the word, but also might learn more about Xapian itself.
- Let us know (either via our bug tracker or by the mailing lists or IRC) if you run into any problems with our documentation, or with bugs you come across while using Xapian.
- Help improve our documentation, either by suggesting changes or by writing up something on our list of missing documentation.

- Contribute new features by working on one of our project ideas
- Tackle an existing bug or feature request, or something you yourself want to see in Xapian.
- Help others get their changes into shape for inclusion in Xapian.

One of the things this guide will do is take you through the process of getting comfortable with the Xapian codebase and submitting your first "patch". There's also lots of more detailed information if you want to get more deeply involved in writing code for Xapian.

2.3.1 Licensing your contributions

If you want a patch to be considered for inclusion in Xapian, you must own the copyright on your changes. Employers often claim copyright on code written by their employees (even if the code is written in their spare time), so please check with your employer if this applies. Be aware that even if you are a student your university may try and claim some rights on code which you write.

Patches which are submitted to Xapian will only be included if the copyright holder(s) dual-license them under each of the following licences:

- GPL version 2 and all later versions (see the file COPYING in `xapian-core` for details).
- MIT/X license:

```
Copyright (c) <year> <copyright holders>

Permission is hereby granted, free of charge, to any person obtaining a copy
of this software and associated documentation files (the "Software"), to
deal in the Software without restriction, including without limitation the
rights to use, copy, modify, merge, publish, distribute, sublicense, and/or
sell copies of the Software, and to permit persons to whom the Software is
furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in
all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING
FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS
IN THE SOFTWARE.
```

The current distribution of Xapian contains many files which are only licensed under the GPL, but we are working towards being able to distribute Xapian under a more permissive license, and are not willing to accept patches which we will have to rewrite before this can happen.

2.3.2 Advice for new contributors

New to Xapian (or even open source)? Don't worry! Here we try to guide you through your first contribution, but if anything is unclear or you want to ask a question, please *get in touch*. This may look a bit daunting the first time, but we're here to help, and a lot of the details will become natural over time.

Checking out and building Xapian

A good way to start learning about Xapian is to check out the code, and get it to build. It's better to use the latest code from the repository rather than a release, as that's what we want the projects to be based on.

We recommend you use Linux or another UNIX-like system for development work, as we're better set up for development on such platforms. In particular we use them ourselves, so can more easily help with any set up issues you may encounter. If you want to run Linux (perhaps virtualised) for development and have no existing preference, we suggest Debian or Ubuntu as our documentation covers these well.

It can take a while to get the code if your network connection is slow, and it may take a while to build and run the testsuite if your computer is slow - while you are waiting, you might want to make a start on the next section.

If you haven't used git before, or want a refresher on "branches", "remotes" and so forth, then this article by James Aylett was written for our students a number of years ago, and may be helpful. There are also a number of free books and resources online, such as Pro Git.

Learn about Xapian's API

It's a good idea to get familiar with Xapian by going through the user guide. The online version has examples in Python, but you can also grab the source and build for other languages; most example code is also available in C++ and PHP.

For more details on individual classes, you may want to look at the automatically generated API documentation. If you're building from git, this will be built for you in `xapian-core/docs`; the API may have some changes between the stable release documented on the website and the latest version in git.

Get familiar with the code

If you're going to be writing code, it's a good idea to read some of Xapian's existing sources, particularly in the main library (`xapian-core`). When you come to write your own, you'll want to follow the style of how the current code works, both in terms of layout (where spaces go and so on) and *how we use various C++ language features*.

Picking something to start with

It can be difficult sometimes to find a place to start, so here are some suggestions:

- Start small.

  By picking a small contribution first, other people can help you with details of how Xapian's documentation, code and so on work. It's much easier to get feedback on a small change to start off with.

  If documentation is your thing, then you might like to take a look at our list of missing documentation.

  On the features side, our bite-sized projects are intended to be suitable for someone new to Xapian to pick up. We've tried to have a range of things to work on across different parts of Xapian.

  Note: No change is too small to consider! Some people's first contributions to an open source project are fixing a single stray letter in documentation, or adding a line break where one is needed.

- Pick something you care about, or which you already have some knowledge of.

  For instance, if you've been using Omega, you might want to pick up a small bug or feature for that. Or if you've studied (or are studying now!) different Information Retrieval weighting schemes, you might want to implement one of the ones we don't currently support.

- Don't be afraid to *ask for help*.

  Please pop onto the mailing list or IRC if you need any help in getting started or picking something to work on.

Do some work

Add or correct some documentation, fix a bug or implement a new feature! You'll probably find our *suggested workflow* helpful. We also have *detailed information* on contributing your changes back to Xapian for inclusion in future releases.

2.3.3 A helpful workflow

If you're working on a bug or new feature, you'll follow something like the following steps. If you haven't worked on Xapian before, it's a good idea to adopt this workflow.

Claim the ticket in trac

Trac is where we keep track of proposed features and known bugs. You'll have to sign up for an account, including verifying your email address, in order to make changes. (This is to prevent unwanted spam on the wiki and in tickets.)

To 'claim' the ticket so others know you're working on it, you can either reassign the ticket to yourself and then claim it, or you can just add a comment saying you're working on it. If you haven't already done so, now's a good time to drop a message either in IRC or to the mailing list saying what you'l be working on.

Note: If there isn't a ticket for what you want to work on, you should create one. (If you're creating one from something on our project list on the wiki then it's good practice to update the wiki to link to the new ticket, so other people can find it easily.)

Create a branch in your local git repository

You want somewhere where you can keep your changes until they're ready, and that won't get confused with anyone else's work. In git these are called branches.

Before you create your branch, you will generally want to make sure your local copy of the code is up to date with the central repository. Generally you can this as follows:

```
$ git checkout master
$ git pull origin/master
```

You can check create your new branch:

```
$ git checkout -b feature-x
```

The branch name doesn't really matter, but you'll probably find it easiest to name it something related to the work that you're doing.

Write a plan

Even the smallest contribution is worth thinking about before starting to type, and with larger changes it's all but essential. If you put your plan in the ticket on trac, it will help when someone else comes to review your patch. Also, if you ask for help, having a plan that someone else can refer to can let them see how you're thinking, so they can provide some useful advice or recommendations more easily.

For a larger piece of work, you ideally want to be able to break down the work into smaller "sub-projects" which can be completed, reviewed and released in turn. (If you're familiar with agile development, think of it as a series of development sprints.)

When planning work on a bug or new feature, you should bear in mind that there are a number of *standards we work to* when accepting changes into Xapian, and the more of them you cover yourself the easier it will be to get your changes into a future release. In particular, it's worth thinking about documentation and tests in advance.

About the only time you don't need to write a plan is when you're making a small change to some documentation, correcting a spelling mistake or making something clearer.

Make your changes!

Now you can start making changes. There's a host of *other information that can help you* if you're writing code. As usual, there are *other people in the community who can help* if you need.

If you discover partway through that your plan isn't working, it's a good idea to stop and write a new plan. You'll have learned something about the problem that you didn't know when you wrote the initial plan, so taking the time to think things through from the beginning can often unblock you and let you start moving again. Of course, if it doesn't clear things up, that's a good time to ask the community for help.

Make commits out of your changes

This is where you probably want to know a little more about git. A very quick introduction is that you first "stage" changes, then you "commit" those changes. You don't have to stage all your changes at once, which means you can keep small notes or parts of future work lying around while you're creating your commits, without them creeping into your commits and confusing matters.

To stage changes for your next commit:

```
$ git add -p
```

The `-p` tells git that you want it to find all the changes, then one by one ask you if you want each staged. Just type `y` to stage a change (it calls them "hunks"), or `n` to skip it this time round. If the file is completely new, you can run `git add <path>` to stage the whole file. (There are lots of other options available in `git add -p`; if you type `?` then it will explain what they all do.)

Then to make a commit:

```
$ git commit -v
```

git will open your editor for you to write a commit message. The `-v` means that your changes will be shown at the bottom of the editor (although they won't be included in the commit message), which helps you do a final check that you're committing only what you want, and everything that is needed.

A good commit in git relies on getting two things right: changes that do a single thing, and a commit message that describes the thing clearly. We have some quick tips on each.

Good commits

Structuring your changes into commits can take a bit of getting used to, but makes it a lot easier for other people to review, both before we merge into Xapian and in the future when someone – which might be you! – needs to understand why a change was made in the past, to help them do whatever work they need to do. There's a good article by Anna Shipman that may help you think about structuring your changes into a set of commits that are easy for others to read.

- Only make *one change* per commit, and make the *whole change* in that commit – you don't want to end up with essential bits of code in a different commit.

  Many people struggle with this at first, and it can be difficult to get into the habit of thinking in terms of the distinct changes to the system rather than in terms of how you did the work. *A plan* here can help structure your commits once you've finished working.

  One of the reasons we suggest using `git add -p` is that it enables you to review every single change that goes into a commit, which can help you put only the right things into it.

- Avoid committing code that has been commented out. If we need it again, it's in the git history.

Good commit messages

Writing a great commit message is important both for people reviewing your code now to help get it ready for a future Xapian release, and for when someone needs to understand how and why a particular change was made, months or years in the future – when that someone might be you!

- Start with a short (50 characters) summary line.

  git (and github) are designed to work better this way. The summary should be in the imperative ("Fix bug on OS X" rather than "Fixed bug on OS X"). This matches git's automatic messages around merges, reverts and so on.

- Follow that with more detail as needed, wrapping long lines at 72 characters (one exception is that long URLs are best not wrapped).

- Describe the effect, not the code. The important thing is for people to be able to read the commit message and understand what you were trying to achieve when you made those changes. That way, if someone needs to work on that part of the code in future, they can understand the purpose of it, and not accidentally remove some useful functionality. (Tests help here, but the commit message is very important.)

There are a few articles around on writing good commit messages; Thoughtbot's "5 Useful Tips For A Better Commit Message" has some good advice.

> **Warning:** Lots of online git tutorials will tell you to write commit messages on the command line, using `git commit -m <message>`. If you do that, you'll never write really good commit messages.

For more details on using git, there are free books and resources online, such as Pro Git.

Contribute your changes

We have *detailed information* to help you here.

2.3.4 Contributing changes

Some things that we look for

Beyond looking for changes that improve Xapian, code that works and so forth, there are a number of things that we aim for when accepting changes. This then is a list of good practices when contributing changes.

Code that compiles cleanly and looks like existing code

We like Xapian to compile without any warnings. In "maintainer mode", which will be how you're building Xapian if you're working from a git clone, all warnings will actually become errors. You should fix the problems rather than change the compilation settings to ignore these warnings.

Please configure your editor to:

- indent each block of code 4 columns from the containing block
- display the tab character by advancing to the next column that is a multiple of 8
- "tab fill" indents: all indents should start with as many tab characters as possible followed by as few spaces as possible
- use Unix line endings (so each line ends with just LF, rather than CR+LF)

We don't currently have a formal coding standards document, so you should try to follow the style of the existing code. In particular, it's a good idea to pay close attention to code alignment and where we have spaces.

Updated documentation

If you add a new feature, please ensure that you've documented it. Don't worry too much about the language you use, or if English isn't your first language. Others can help get the documentation into shape, but having a first draft from the person who wrote the feature is usually the best way to get started.

- API classes, methods, functions, and types should be documented by documentation comments alongside the declaration in `include/xapian/*.h`.

  These are collated by doxygen – see doxygen's documentation for details of the supported syntax. We've decided to prefer to use `@` rather than `\` to introduce doxygen commands (the choice is essentially arbitrary, but `\` introduces C/C++ escape sequences so `@` is likely to make for easier to read mark up for C/C++ coders).

- The documentation comments don't give users a good overview, so we also need documentation which gives a good overview of how to achieve particular tasks.

  If there's relevant documentation already in the user guide, then you should update that. For completely new features, you should create either a "how to" or an "advanced feature" document in the user manual, so that people can get started without having to start with the API documentation.

- Internal classes, etc should also be documented by documentation comments where they are declared.

Automated tests

If you're fixing a bug, you should first write a regression test. The test will fail on the existing code, then when you fix the bug it will pass. In the future, the test will make sure no one accidentally re-introduces the same bug.

If you're adding a new feature, you'll want to write tests that it behaves correctly. Thinking about the tests you need to write can often help you plan how to implement the feature; it can also help when thinking about what API any new classes or methods should expose.

Updated attributions

If necessary, modify the copyright statement at the top of any files you've altered. If there is no copyright statement, you may add one (there are a couple of Makefile.am's and similar that don't have copyright statements; anything that small doesn't really need one anyway, so it's a judgement call). If you've added files which you've written from scratch, they should include the GPL boilerplate with your name only.

If you're not in there already, add yourself to the `xapian-core/AUTHORS` file.

Consider backporting bug fixes

If there's an active release branch, please check if the bug is present in that branch, and if the fix is appropriate to backport - if the fix breaks ABI compatibility or is very invasive, you may need to fix it in a different way for the release branch, or decide not to backport the fix.

License grant

We ask everyone contributing changes to Xapian to *dual-license* under the GPL (which Xapian currently uses) and the MIT/X license (which we would like to move to in future). The simplest way to do this is to drop an email to the xapian-devel mailing list stating that you own the copyright on your changes and are happy to dual-license accordingly.

Submit your patch

There are two ways of working, depending on whether you want to use Github or not. In both cases, review and acceptance of the changes will generally go more easily if you've included tests, updated documentation and so on *as discussed earlier*.

Attach a patch directly to the trac ticket

We find patches in unified diff format easiest to work with. `git diff` produces the right output for a single commit (or `git format-patch` for a series of commits).

Someone from the community will then be able to review the patch and decide if it needs further work before integrating. If so, they'll leave comments on the trac ticket (trac will generally email you if you're marked as the owner, or you can explicitly add yourself to the "cc" list for a ticket).

Open a Pull Request on github

Github pull requests provide a web-based interface for review and discussion of changes before they are accepted into Xapian. Github's documentation explains how you can go about opening them.

If your patch is a sub-project in a larger piece of work, then it's important not to assume the patch is fine as it stands and to immediately start the next sub-project. Instead you should concentrate on completing the sub-project before moving on. Since you'll almost always have to wait at least a little time to get feedback on any changes, you may want to put the code and tests up while still working on documentation.

You should add further changes to pull requests by creating additional commits locally, typically by using `git commit --fixup`, and then pushing the branch up to Github. Only once everything's been approved should you squash your commits together to keep the history clean.

Note: Once you've opened a pull request, you shouldn't have to close it until it's merged (in which case we'll generally close it for you). Even if you need to redo some work, you can either add fixup commits or (with agreement from whoever is reviewing the PR) unwind your work and create completely new commits, force pushing to replace the previous commits in the pull request.

It makes it much harder to review if you close a pull request in the middle of a review only to open another with similar code.

2.3.5 Handy tips for aiding development

Disabling documentation builds

If you are find you are repeatedly changing the API headers (in include/) during development, then you may become annoyed that the docs/ subdirectory will rebuild the doxygen documentation every time you run "make" since this takes a while. You can disable this temporarily (if you're using GNU make), by creating a file "docs/GNUmakefile" containing these two lines:

```
%:
    @echo "Skipping 'make $@' in docs"
```

Note that the whitespace at the start of the second line needs to be a single "tab" character!

Don't forget to remove (or rename) this and check the documentation builds before committing or generating a patch though!

Integration syntax checking with your editor

If you are using an editor or other tool capable of running syntax checks as you work there you can use the *make* target 'check-syntax'. For 'emacs' users this works well with 'flymake'. Usage from a shell:

```
make check-syntax check_sources=api/omdatabase.cc
```

2.4 Mentoring new contributors

Xapian frequently participates in Google Summer of Code (GSoC), which encourages student developers to contribute to open source projects. We also welcome new contributors at any time.

This section contains advice and information for anyone within the community helping newcomers come up to speed as members of our community. Especially during GSoC, we welcome anyone to act as a mentor who is prepared to commit enough time. Many of our GSoC students have gone on to mentor in subsequent years.

Note: If you're looking for help in getting started, then we have a guide for potential GSoC students (which is worth reading through even if you aren't participating in GSoC). The *Contributing to Xapian* section of this guide also has useful pointers.

2.4.1 There's plenty of help

You may be daunted by the idea of mentoring someone else, particularly if you only have a limited amount of expeirence with Xapian itself. Please don't let this put you off! Those who have been around in the community for longer are generally happy to provide advice and assistance, and there's also a mentor guide written for GSoC (including contributions from Olly Betts from Xapian).

As with almost everything in life, **it's good to ask for help**. Xapian as a community has years of experience with Summer of Code, and some individuals have mentored five or more times. If we can't figure out a problem together as a community, we can also ask for help from other projects and mentors, and from the Google Summer of Code team themselves.

2.4.2 Helping newcomers step by step

Building Xapian

The first step should always to ensure that a new contributor can build Xapian on a machine they have access to. Our *getting started information* is a good guide to follow.

Most Xapian developers use one or more of Debian, Ubuntu, and macOS. While it's possible to develop for Xapian on a wide range of operating systems, if someone runs into problems with something else it's often easier for them to work with a linux virtual machine running on their computer. As the getting started information says, we recommend using Virtual Box and the latest LTS (long-term support) release of Ubuntu.

Note: If a new contributor runs into problems, it's worth getting them to explicitly confirm exactly what steps they've taken. It's easy to skip a step, or to try something else when you're working through issues – but when it comes to helping someone else, you need to be sure you know exactly what they've done.

A lot of our project communication happens on IRC, and it's difficult to read long outputs of commands such as `make` there. It's helpful to have people copy anything substantial into something like Pastebin or Gist and then provide a link in IRC. This can be used for command output, or for the contents of files such as `config.log`.

Getting familiar with the API

While the urge to jump right in and start fixing bugs or adding features is often strong, it's a good idea for a new contributor to become familiar with Xapian's API early on. This is particularly true for Summer of Code students, who often won't have used Xapian previously.

The Xapian user manual covers the concepts behind Xapian, works through a practical example of indexing and searching documents using Xapian, and also covers a range of more advanced features. The online version uses python, but you can grab the source code and build for a range of languages, including C++.

Note: Almost every contributor should get familiar with Xapian's C++ API. Anyone who's adding new APIs to Xapian should also think about how that API will be used from bindings languages, so it's often helpful for them to have at least used Xapian via python or one of the other languages we support.

Completing a small task

We ask all Summer of Code students to complete at least one small task, effectively as part of their application. This could be fixing a bug, tidying up some code, improving test coverage, or completing a small feature.

A main part of the reason for this is to help new contributors get used to *the way we manage changes* to Xapian. It's good to pick something small, and go through the entire process of planning and doing the work, creating a pull request, and getting everything merged. Firstly, that means that they've become a contributor to Xapian before Summer of Code even begins. Secondly, it means that subsequent contributions as part of the GSoC project should be smoother. Finally, it helps someone new to open source and collaborative development to start thinking in terms of small changes, merged quickly.

Most students will have an idea of what project they are interested in, from our list of project ideas. Some of those will have a small first step, or sometimes a bug or similar in the general area of the codebase of the project. For everyone else, we keep a list of bite sized projects, and there are also some bugs in our tracker marked as suitable for newcomers to tackle.

Helping them through their first contribution

Just as with getting Xapian built for the first time, new contributors may need support in getting through our contribution flow. There are some common things to watch out for.

Pull requests where the automated tests aren't passing

> This includes a special test run which checks that the code diff follows some of our conventions, such as around spaces and blank lines. Generally, the error messages should help track down the problem.
>
> It's possible (and a good idea) to run all the tests locally before opening a pull request. The code diff checks can be run by piping the output of `git diff` through the `xapian-maintainer-tools/xapian-check-patch` script. Something like this is often what you want:
>
> (The `git diff` command there will output the changes in your local commits compared to the "master" branch.)

Not following our *coding conventions*

> We can't automate checks for all of these, but we also don't expect anyone to be able to spot all possible problems. One of the reasons pull request reviews are open is so that several different people can help spot and straighten out issues, and get a contribution over the line.

Not following our conventions for *pull request flow*

> In particular, first-time contributors often need reminding not to force-push branches once a PR is open. It feels tidier to have a tidy list of commits. However, it makes it harder for reviewers to check that earlier comments have been addressed. Contributors should use "fixup" commits, as described in our documentation, and only tidy up commits right before a pull request is merged.
>
> A similar problem is a pull request with lots of commits without good commit messages. Each commit in a pull request should make a single, well-described change, including any necessary tests and documentation.

Some of these take a long time to get used to, and even experienced developers will make mistakes. That means that it's worth checking for the basics on every pull request.

2.4.3 Expectations of mentors

We operate a "group mentoring" approach, which means you can – and should! – help any students you can during the summer. Where possible, we expect mentors to find time every week to engage with Xapian and our Summer of Code students. Here are some ways to do that.

Answer questions on the mailing list and IRC

> It's demotivating to ask a question and get no reply. Sometimes even just a response that says you don't know can help reassure a new contributor that they aren't on their own.
>
> Particularly during the early phases of Summer of Code, there are a lot of questions that come up repeatedly. New contributors regularly need help getting Xapian built and installed on their computers. People often need pointing at our guidance for potential students (the GSoC site sends people straight to our ideas list, and it's easy to miss the links we provide to furhter information). So something as simple as chipping in to point people to existing information and documentation can be incredibly valuable.

Help review pull requests

> The core of a contribution to Xapian is often a pull request. Before it's merged by one of the Xapian team, we want to make sure it's in *good shape*. Anyone can check over our guidelines on what we're looking for, and provide feedback to a contributor on how to improve their pull request.

Provide feedback on design ideas

Most Summer of Code projects have a knotty or interesting problem at the heart of them. That's what makes them appealing to work on over a period of months. However, that means that there's often one or more points during the project where some decisions have to be made. APIs need designing, data structures need choosing, and sometimes different competing algorithms need assessing.

While we expect our students to do most of the work here, getting timely feedback and input from the rest of the community is often important in keeping a project on track. As with reviewing pull requests, anyone can look over a proposal and provide their thoughts.

Note: API design is a particularly difficult problem, and we generally do not expect any one person (student or not!) to design a great API on their own. It's not always obvious the best approach until you've written code that uses an API in a range of different situations.

We generally recommend that projects that require a new API start by implementing a very simple one. Ideally this will leave time later in the project to revise the initial version based on feedback and experience of actually using it. (Summer of Code students also include a section on possible future improvements in their project write-up. If there isn't enough time to improve an API based on feedback, that can always become a future project!)

Encourage small, regular contributions that can be merged

We recommend structuring any project as a series of small sub-projects, each of which can be submitted as a pull request, reviewed, and merged. It's usually possible to start work on the next sub-project while the previous one is going through review.

As well as encouraging contributors to submit small changes, it's also important that they address review comments quickly. It's all too easy to move on to the next sub-project, but never actually get the previous one merged. It's far better to spend time getting two or three sub-projects merged than have pull requests for four or five none of which is in a good enough state to be merged.

Our best Summer of Code students have often had early contributions released during their project. Our experience shows that contributors are more likely to become longer-term members of the Xapian community if their early work can be merged and released.

As well as group mentoring, every student has a specific mentor assigned, who is there to make sure there is always someone looking out for them. You should keep in regular contact with the student you're assigned to, making sure that they're getting the support they need from the community.

Note: Mentors, as well as students, can have something unexpected come up during the summer. Plans change, work becomes busier, or any one of a hundred things can mean you suddenly have less time than you anticipated.

If something comes up, please let Xapian's "org admins" for Summer of Code know as soon as possible. Our group mentoring approach makes it easier to cope with people who have to step away from Summer of Code, but we need to know to ensure we can support all our students as best we can.

2.5 License

This license applies to all documentation and example code in this book.