Automating Compliance: Linux Case Study for Solving the Problem at the “Source”

Kate Stewart, Philippe Ombredanne & Greg Kroah-Hartman
Automating Compliance

Challenge:
Accurate Recognition

Licenses must be able to be accurately communicated, and obligations associated with the licenses clearly understood.

Source: https://athithanacademy.com/2017/03/17/code-review/
Automating Compliance

Challenge:

Developers != Lawyers

We need to make it easy for developers to indicate the licensing without boiler plate that changes over time.

Source: https://athithanacademy.com/2017/03/17/code-review/
Scanning tools

Quick and dirty; no tooling needed

Look for relevant words / fragments:
- “licen”
- “redist”
- “copyright”
- common license fragments: “bsd”, “gpl”, “general public”, “cddl”, ...

Manual searches

`grep -nri`
(or your favorite command line args)

`Ctrl-F`
(or your favorite editor’s equivalent)
FOSSology is used to scan a codebase for licenses

Performs textual analysis and regular expression scanning to identify likely license notices and references

Supplemented with manual review to remove false positives and investigate unusual findings

https://github.com/fossology/fossology
Scanning tools…

ScanCode scans code for license, copyright, package manifests and dependencies and other interesting and binary code files.

ScanCode Toolkit by nexB

https://github.com/nexB/scancode-toolkit
Various other scanning tools and services, including open source and proprietary / commercial options

Some include security vulnerability detection

Some include initial free tiers for open source projects (read carefully how they define “open source” and “projects”)
Why now?

Developers like simple.

Adoption of SPDX short form license identifiers by projects, is enabling the accurate identification of licenses to be done by simpler tooling, rather than complicated heuristics & manual checking.

Licensing has evolved to be per file (and sometimes even per snippet), so only looking at the project level licensing no longer suffices.
Communicating License Information

From the SPDX License List:

“...a list of commonly found licenses and exceptions used in free and open source and other collaborative software or documentation.”

“The purpose of the SPDX License List is to enable easy and efficient identification of such licenses and exceptions in an SPDX document, in source files or elsewhere.”

SPDX License List

https://spdx.org/licenses
Communicating License Information

From the License List:

“...a list of commonly found licenses and exceptions used in free and open source software or other collaborative software or documentation.

“The purpose of the SPDX License List is to enable easy and efficient identification of licenses and exceptions in an SPDX document, in source files or elsewhere.”

Examples:

- BSD-2-Clause
- BSD-3-Clause
- GPL-3.0-only
- GPL-3.0-or-later
- MIT
- MPL-2.0

https://spdx.org/licenses
One-line comment in each source code file to unambiguously designate the applicable license(s)

Examples:

// SPDX-License-Identifier: Apache-2.0

// SPDX-License-Identifier: GPL-2.0-only OR MIT

// SPDX-License-Identifier: Apache-2.0 AND MIT

SPDX Short-Form IDs

Usage example:
The REUSE Initiative (from Free Software Foundation Europe) provides **best practices** in communicating license information for an entire package, and **tools** to assist in confirming compliance with those practices.

Makes use of SPDX short-form identifiers

Includes recommendations for how and where to place copyright notices, license references and license texts, version 2.0 incorporated feedback from Linux Kernel work.

REUSE Initiative

https://reuse.software

The REUSE website and logo are copyright © FSFE e.V. The REUSE logo is licensed under Creative Commons Attribution-ShareAlike 4.0.
No Licensing Bugs!

If you can’t automatically detect the license accurately with tools, it is a source code bug!

Source: https://www.bedbugs.org/pictures/
Case Study: Linux Kernel

Source: http://fcgp.sourceforge.net/lgp/zoom.jpg
The Linux Kernel changes rapidly...

Files:
• 59,801 (4.12)
• 60,538 (4.13)
• 61,258 (4.14)
• 62,271 (4.15)
• ...

Overall: about 8.5 changes per hour*

26 years of code incorporated and shared with other projects.

* Based on 2017 Linux Kernel Report & https://github.com/gregkh/kernel-history
The Linux Kernel is GPL version 2 only!
Which version of GPL version 2?

- There are 6 FSF created variants of GPL version 2 license text.

- COPYING file is GPL version 2 – “mark 4”
  ➔ multiple other variants are present in kernel source files.

- COPYING indicates Linux User Space API files need special handling
  ➔ Required new SPDX exception “Linux-syscall-note” to be created.

The Linux Kernel is GPL version 2 only… well mostly.

- User Space API files
- Imported and shared code with other projects
- Dual licensed code
- Build tools
- …
Source Code Scanner Results:

- **Linux 3.16** (2014-8-3)
  - Debian Jessie kernel, FOSSology & ScanCode analysis
  - 940+ different license texts detected by scanners

- **Linux 4.13** (2017-9-3)
  - 1020+ different license texts detected by ScanCode
  - About 700+ are variants on simple GPL declaration
  - 11,000+ files without any license in the file
What Source Code Scanners see:

**Obfuscated**: Seen in a GPL-shy external kernel module
- `MODULE_LICENSE("\x47\x50\x4c\x20\x76\x32");`
- (e.g. GPL v2 in ASCII)

**Terse**: In `drivers/mtd/chips/cfi_cmdset_0001.c`:
- (C) 2000 Red Hat. GPL'd

**Hot**: In the kernel `drivers/wl/slaves/wl_therm.c` thermal drivers code:
- This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License...
Need Accurate at File Level

Methodology: Linux Kernel 4.13 Analyzed at File Level
- 2 scanners (Windriver, ScanCode) used to generate initial SPDX files
- Diff’d the SPDX documents and generated 60,538 row spreadsheet
- Determined license governing each file
  - When 2 scanners agreed on license, determined that was license
  - When no detectable license, COPYING license determined to apply
  - When conflict, manually inspect and resolve.
  - Some ambiguous cases, deferred to study further in future.
- Cross checked with 3rd scanner output (FOSSology 3.1).

Effort: 80 logged hours of manual analysis
Detected Licenses in 4.13

Preferred (12):
- GPL-2.0, GPL-2.0+, GPL-1.0, GPL-1.0+, LGPL-2.0, LGPL-2.0+, LGPL-2.1, LGPL-2.1+, MIT, BSD-2-Clause, BSD-3-Clause, BSD-3-Clause-Clear.

Other (14):
- Apache-2.0, GPL-3.0, ISC, GFDL-1.1, Artistic-2.0, MPL-1.1, MPL-2.0, X11, OpenSSL, CC0, Unlicense, Zlib

Exceptions (4):
- Linux-syscall-note, GCC-exception-2.0, Bison-exception-2.2, mif-exception

86 combinations detected
- described as SPDX License Expressions using “AND”, “OR”, “WITH” & “(, )”.

THE LINUX FOUNDATION
### 4.13 File Level Licensing

<table>
<thead>
<tr>
<th>License</th>
<th># Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPL-2.0-only</td>
<td>31,449</td>
</tr>
<tr>
<td>GPL-2.0-or-later</td>
<td>10,750</td>
</tr>
<tr>
<td>MIT</td>
<td>2,055</td>
</tr>
<tr>
<td>GPL-2.0-only WITH Linux-syscall-note</td>
<td>1,221</td>
</tr>
<tr>
<td>GPL-1.0-or-later</td>
<td>1,157</td>
</tr>
<tr>
<td>GPL-2.0-only OR BSD-2-Clause</td>
<td>832</td>
</tr>
<tr>
<td>GPL-2.0-only OR BSD-3-Clause</td>
<td>826</td>
</tr>
<tr>
<td>ISC</td>
<td>439</td>
</tr>
<tr>
<td>GPL-2.0-or-later OR MIT</td>
<td>436</td>
</tr>
<tr>
<td>GPL-2.0-or-later OR BSD-3-Clause</td>
<td>371</td>
</tr>
</tbody>
</table>

- Total # Files Analyzed: **60,538**
- NO SPDX tags needed: **9,835**
  - Documentation, etc.
- Missing licensing: **11,139**
  - Files with no detectable license default to COPYING
Handling System Calls

- The Linux Foundation and key kernel developers worked with the SPDX legal team over the summer to define:
  - Linux-syscall-note

- Use "WITH" operator in SPDX expressions

[Link to SPDX Note: https://spdx.org/licenses/Linux-syscall-note.html]
Linux 4.14

Update for 4.14 was based on the analysis from 4.13 and additions:

- Files with no detectable license: Add SPDX identifiers based on COPYING to all files with no detectable license references in them (11,139 files)
- Add “WITH Linux-syscall-note” to all files using GPL/LGPL licensing with User Space API (/uapi/) files (1,452 files).

At Maintainers Summit, Oct. 2017, kernel developers came to agreement on including SPDX-License-Identifier in source files.

See:

https://lwn.net/Articles/738235/
https://lwn.net/Articles/739183/

<table>
<thead>
<tr>
<th>SPDX-License-Identifier</th>
<th># files</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPL-2.0 WITH Linux-syscall-note</td>
<td>1200</td>
</tr>
<tr>
<td>GPL-2.0+ WITH Linux-syscall-note</td>
<td>169</td>
</tr>
<tr>
<td>((GPL-2.0 WITH Linux-syscall-note) OR BSD-2-Clause)</td>
<td>21</td>
</tr>
<tr>
<td>((GPL-2.0 WITH Linux-syscall-note) OR BSD-3-Clause)</td>
<td>17</td>
</tr>
<tr>
<td>LGPL-2.1+ WITH Linux-syscall-note</td>
<td>15</td>
</tr>
<tr>
<td>GPL-1.0+ WITH Linux-syscall-note</td>
<td>14</td>
</tr>
<tr>
<td>((GPL-2.0+ WITH Linux-syscall-note) OR BSD-3-Clause)</td>
<td>5</td>
</tr>
<tr>
<td>LGPL-2.0+ WITH Linux-syscall-note</td>
<td>4</td>
</tr>
<tr>
<td>LGPL-2.1 WITH Linux-syscall-note</td>
<td>3</td>
</tr>
<tr>
<td>((GPL-2.0 WITH Linux-syscall-note) OR MIT)</td>
<td>3</td>
</tr>
<tr>
<td>(GPL-2.0 WITH Linux-syscall-note) AND MIT</td>
<td>1</td>
</tr>
</tbody>
</table>
Guidelines for adding SPDX identifiers

<table>
<thead>
<tr>
<th>File Type</th>
<th>SPDX License tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>C source:</td>
<td>// SPDX-License-Identifier:</td>
</tr>
<tr>
<td>C header:</td>
<td>/* SPDX-License-Identifier: */</td>
</tr>
<tr>
<td>ASM:</td>
<td>/* SPDX-License-Identifier: */</td>
</tr>
<tr>
<td>scripts:</td>
<td># SPDX-License-Identifier:</td>
</tr>
<tr>
<td>.rst:</td>
<td>.. SPDX-License-Identifier:</td>
</tr>
<tr>
<td>.dts{i}:</td>
<td>// SPDX-License-Identifier:</td>
</tr>
</tbody>
</table>


**NOTE:**

// on first line has been determined by Linus* as preference. This is probably source of largest number of developer comments so far.

* “So I want the format to be _fixed_. On the very first line that the file format allows, with no whitespace garbage, and no "let's try to make it pretty and match the rest". [https://lkml.org/lkml/2017/11/25/125](https://lkml.org/lkml/2017/11/25/125)
"grep" Goal

Accurate licensing should be able to be determined by:

```bash
git grep -h SPDX-License-Identifier: | sort | uniq --count
```

It should not require **80+ hours** to manually look through files, after scanners (with complicated heuristics/machine learning) to make judgements about what is the actual intention, then ensure all licenses detected (in addition to GPL-2.0) have their terms respected.

* 3 data points indicate that this is about what it takes to properly clear a new version of the Linux kernel, and generate the appropriate artifacts.
Progress towards goal...
Next Steps: help wanted!

- **Support** `checkpatch.pl` **update** to prevent intro of new files without SPDX identifiers
- Work with maintainers to get all subsystems to use SPDX identifiers
- Help to submit patches to get more files tagged with SPDX identifiers