A Community Take on the License Compliance Industry

Stefano Zacchiroli

Debian Developer
OSI Board Director

20 March 2016
LibrePlanet 2016
MIT
Cambridge, MA, USA

What is compliance?

"respect the terms of the applicable FOSS license"

Definition (Compliance)

Noun:

1. acting according to certain accepted standards
2. disposition or tendency to yield to the will of others
3. the act of submitting;
   usually surrendering power to another

— WordNet 3.0 (2006)

What is compliance?

“respect the terms of the applicable FOSS license”

Definition (Compliance)

Noun:

1. acting according to certain accepted standards
2. disposition or tendency to yield to the will of others
3. the act of submitting;
   usually surrendering power to another

— WordNet 3.0 (2006)

What is compliance?

“respect the terms of the applicable FOSS license”

Definition (Compliance)

Noun:

1. acting according to certain accepted standards → by whom?
2. disposition or tendency to yield to the will of others → who?
3. the act of submitting;
   usually surrendering power to another → to whom?

— WordNet 3.0 (2006)

Compliance ← seen from “the FOSS community”

Who:

- activists
- developers
- (copyright holders)

Goals:

- pursue a political strategy (e.g., copyleft)
- make sure everyone “play by the rules” (≈ “accepted standards”)
  - the legal soundness of rules is barely relevant here
- (defend an investment)

Compliance ← seen from “the FOSS community”

Who:

- activists
- developers
- (copyright holders)

Goals:

- pursue a political strategy (e.g., copyleft)
- make sure everyone “play by the rules” (≈ “accepted standards”)
    - the legal soundness of rules is barely relevant here
- (defend an investment)

Compliance ← seen from “the software industry”

Who:

- for profit IT vendors who ship FOSS
- in context:
  - links in a long software supply chain
  - market pressure
- FOSS competence: very variable
- commitment to FOSS ideals: free riders ⟷ contributors

Goals:

- minimize legal risks
- minimize compliance cost
- (reassure the next link in the supply chain)

Compliance ← seen from "the software industry"

Who:

- for profit IT vendors who ship FOSS
- in context:
  - links in a long software supply chain
  - market pressure
- FOSS competence: very variable
- commitment to FOSS ideals: free riders ⟷ contributors

Goals:

- minimize legal risks
- minimize compliance cost
- (reassure the next link in the supply chain)

The compliance industry

Customers:

- “the software industry”

Products:

- code scanners
  - provenance tracking
  - linting for common “IP” issues
  - adherence to *ad-hoc* “IP” policy
- Bill Of Material (BOM) reporters
  - long-term software maintenance
- software qualification (e.g., security flaws, quality metrics, etc.)

By-products:

- reports on the state of the FOSS ecosystem

The compliance industry

Customers:

- “the software industry”

Products:

- code scanners
  - provenance tracking
  - linting for common “IP” issues
  - adherence to *ad-hoc* “IP” policy
- Bill Of Material (BOM) reporters
  - long-term software maintenance
- software qualification (e.g., security flaws, quality metrics, etc.)

By-products:

- reports on the state of the FOSS ecosystem

The compliance industry

Customers:

- “the software industry”

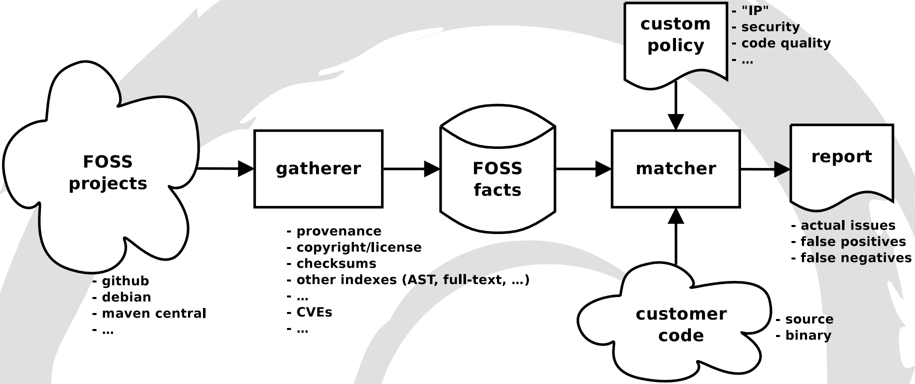
Products:

- code scanners
  - provenance tracking
  - linting for common “IP” issues
  - adherence to *ad-hoc* “IP” policy
- Bill Of Material (BOM) reporters
  - long-term software maintenance
- software qualification (e.g., security flaws, quality metrics, etc.)

By-products:

- reports on the state of the FOSS ecosystem

The compliance toolchain



FOSS projects

- github
- debian
- maven central
- …

gatherer

- provenance
- copyright/license
- checksums
- other indexes (AST, full-text, …)
- …
- CVEs
- …

FOSS facts

custom policy

- "IP"
- security
- code quality
- …

matcher

customer code

- source
- binary

report

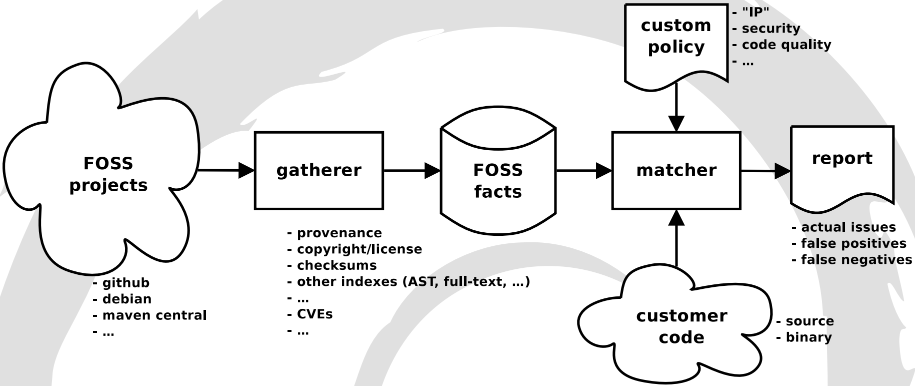
- actual issues
- false positives
- false negatives

A note about copyright-related facts:

✗ ~~*foo.c is under GPL3+*~~

✓ *foo.c, when we last saw it (UNIX timestamp 1454060276) at http://git.example.com/foo/, had SHA1* f1d2df2... *and copyright header “foo is free software [. . .] under the terms of the GNU General Public License, version 3 or any later version”*

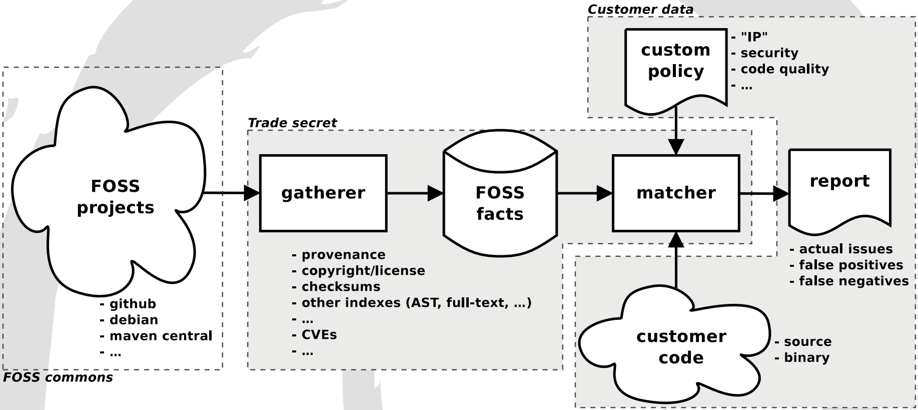
The compliance toolchain



A note about copyright-related facts:

- ✗ ~~*foo.c is under GPL3+*~~
- ✓ *foo.c, when we last saw it (UNIX timestamp 1454060276) at http://git.example.com/foo/, had SHA1 `f1d2df2...` and copyright header "foo is free software [. . .] under the terms of the GNU General Public License, version 3 or any later version"*

The compliance toolchain . . . and the commons



A community critique of today's compliance industry

- the tragedy of the ethical software entrepreneur
  - non-free tooling
  - non-free data
- no (or poor) reproducibility
- who controls the ~~controllers~~ policy makers?
- the cultural impact of by-products

A community critique of today's compliance industry

- the tragedy of the ethical software entrepreneur
  - non-free tooling
  - non-free data
- no (or poor) reproducibility
- who controls the ~~controllers~~ policy makers?
- the cultural impact of by-products

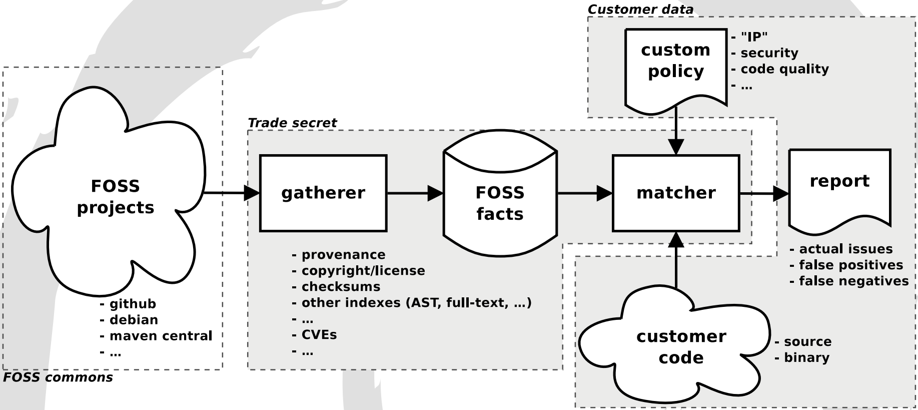
A community critique of today's compliance industry

- the tragedy of the ethical software entrepreneur
  - non-free tooling
  - non-free data
- no (or poor) reproducibility
- who controls the ~~controllers~~ policy makers?
- the cultural impact of by-products

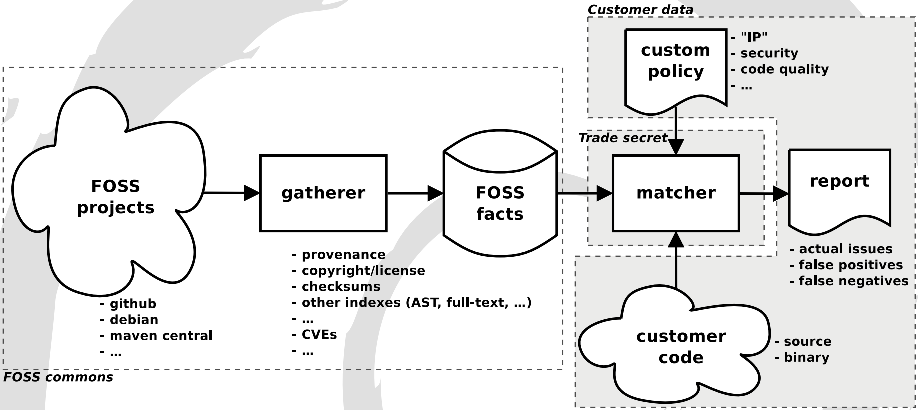
A community critique of today's compliance industry

- the tragedy of the ethical software entrepreneur
  - non-free tooling
  - non-free data
- no (or poor) reproducibility
- who controls the ~~controllers~~ policy makers?
- the cultural impact of by-products

The compliance toolchain — today



The compliance toolchain — tomorrow?



Benefits

for “the FOSS community”:

- free tools and data (e.g., to point allies to)
- independent analyzability of the state of the ecosystem

for “the software industry”:

- reduced risks of vendor lock-in
- more disintermediation, FOSS facts from the source

for the compliance industry:

- community buy in
- effort sharing → reduced costs

Benefits

for “the FOSS community”:

- free tools and data (e.g., to point allies to)
- independent analyzability of the state of the ecosystem

for “the software industry”:

- reduced risks of vendor lock-in
- more disintermediation, FOSS facts from the source

for “the compliance industry”:

- community buy in
- effort sharing → reduced costs

Benefits

for “the FOSS community”:

- free tools and data (e.g., to point allies to)
- independent analyzability of the state of the ecosystem

for “the software industry”:

- reduced risks of vendor lock-in
- more disintermediation, FOSS facts from the source

for the compliance industry:

- community buy in
- effort sharing → reduced costs

A small scale example: Debsources

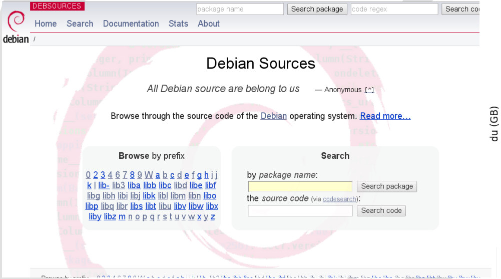
1. an infrastructure to publish Debian source code on the Web
2. a notable instance indexing *all* Debian source code to date: `http://sources.debian.net`

For developers:

- browse/search source code
- syntax highlighting
- pinpoint code lines, annotate

For data miners:

- 20+ years of FOSS history
- live change monitoring
- source code indexing



Debian Sources

All Debian source are belong to us — Anonymous [^]

Browse through the source code of the Debian operating system. Read more…

Browse by prefix

0 2 3 4 6 7 8 9 W a b c d e f g h i j k l lib- lib3 liba libb libc libd libe libf libg libh libi libj libk libl libm libn libo libp libq libr libs libt libu libv libw libx liby libz m n o p q r s t u v w x y z

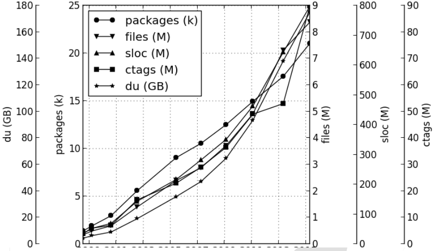
Search

by *package name*:

Search package

the *source code* (via codesearch):

Search code



Debsources — coverage

Covered releases:

- all stable releases from Debian Hamm (1997) to Jessie (2015)
- LTS security updates
- development releases: testing, unstable, experimental, . . .

Update frequency: 4 times a day (at each Debian archive change)

Overall content: (Oct 2015)

- 90 K source packages
- 790 GB of source code
- 45 M source code files
  - 18 M *distinct* SHA256
- 4.3 B lines of code
- 485 M developer-defined symbols (ctags)

more stats at
`http://sources.debian.net/stats/`

Debsources — copyright & license information

Data providers of copyright & license information:

- debian/copyright files
- scans of the Debsources corpus[1]
  - FOSSology[2]
  - Ninka[3]

Data consumers: public, well-documented API

[1] joint work with Daniel M. German and Matthieu Caneill, upcoming publication

[2] http://fossology.org

[3] http://ninka.turingmachine.org

Use case #1: detect bit-identical reuse

Example

```
http://sources.debian.net/api/sha256/?checksum=
ae8e672aaa16bbdf734eabefaf2ee5987013d726868f776de1728f6a36a0ae2d
```

```
{ "count": 3,
  "sha256":
    "ae8e672aaa16bbdf734eabefaf2ee5987013d726868f776de1728f6a36a0ae2d",
  "results": [
    { "path": "coreutils/ls.c",
      "version": "1:1.22.0-12",
      "package": "busybox" },
    { "path": "coreutils/ls.c",
      "version": "1:1.22.0-15",
      "package": "busybox" },
    { "path": "coreutils/ls.c",
      "version": "1:1.22.0-9+deb8u1",
      "package": "busybox" } ] }
```

It is now trivial to develop a source code scanner that uses Debsources as backend to detect bit-identical reuse of source code files present in Debian.

Use case #2: detect reuse with modification

Debsources can support simple fingerprinting techniques:

- ctags searches — *"show me the files that define this function/ variable/ class/ etc."*

Example

`http://sources.debian.net/api/ctag/?ctag=pcre_compile`

```
{ "count": 400,
  "ctag": "pcre_compile",
  "results": [
    { "path": "glib/pcre/pcre_compile.c", "line": 7565,
      "package": "glib2.0", "version": "2.33.12+really2.32.4-5"
    },
    { "path": "pcre_compile.c", "line": 7563,
      "package": "pcre3", "version": "1:8.30-5"
    },
    { "path": "libasync/pcre.c", "line": 4097,
      "package": "mailavenger", "version": "0.8.3rc1-1"
    }, [...] ] }
```

- *ad hoc* regexp searches (powered by codesearch.debian.net)

Use case #3: SPDX generation

When instantiated to a specific source package, machine-readable debian/copyright files can be used to automatically generate SPDX.

Example (SPDX export)

```
http://sourcesdev.debian.net/copyright/license/gnubg/1.04.000-1/
http://sourcesdev.debian.net/copyright/spdx/gnubg/1.04.000-1/
```

```
SPDXVersion: SPDX-2.0
DataLicense:CC0-1.0
DocumentName: GNU Backgammon
FileName: bearoffgammon.h
FileChecksum: SHA256: 4e87bfe929021d710b4046b570b2042489c2cd7cdabc9ea46572b1
LicenseConcluded: GPL-3+
FileCopyrightText: <text>1984, 1989-1990, 1995-1997, 1999-2011
    Free Software Foundation, Inc.
  1996 Claes Thornberg (claest@it.kth.se)
  1998-1999 Mark Spencer <markster@marko.net>
  2000 Jonathan Blandford
[...]
```

Credits: Orestis Ioannou, GSoC 2015. Status: beta, dev. preview

Parting thoughts

A benchmark for the compliance industry:

- exhaustivity → hopeless without a collaborative effort!
- quality → composes well; industry-private + open DBs?
- granularity → lot to do, great competition playground!

Parting thoughts

A benchmark for the compliance industry:

- exhaustivity → hopeless without a collaborative effort
- quality → composes well; industry-private + open DBs?
- granularity → lot to do, great competition playground!

Parting thoughts

A benchmark for the compliance industry:

- exhaustivity → hopeless without a collaborative effort
- quality → composes well; industry-private + open DBs?
- granularity → lot to do, great competition playground!

Parting thoughts

A benchmark for the compliance industry:

- exhaustivity → hopeless without a collaborative effort
- quality → composes well; industry-private + open DBs?
- granularity → lot to do, great competition playground!

- the compliance industry has potential to help well-meaning actors
- but to do so properly we need free tools and data
- an open compliance DB might benefit everybody

Thanks!

Questions?

Stefano Zacchiroli
zack@debian.org
http://upsilon.cc/zack

about the slides:
available at https://upsilon.cc/~zack/talks/2016/2016-03-19-libreplanet-compliance.pdf