Debsources as a Community Curated DB of Copyright and License Information

Stefano Zacchiroli

Debian Developer
Former Debian Project Leader
OSI Board Director

8 October 2015
Supply Chain Mini Summit
LinuxCon Europe
Dublin, Ireland

Outline

1. Debian and compliance
2. Debsources
3. Debsources for compliance

Outline

1. Debian and compliance

2. Debsources

3. Debsources for compliance

Debian — the operating system

http://www.debian.org

flagship product: Debian stable

- binary distribution
- released every ≈24 months
- 10 hw architectures
- 5 years LTS security support

in the Jessie release:

- 43’000 binary packages
- 800M lines of code

Debian is often credited as the largest curated Free Software collection

- base for ≈140 (48%) distributions — DistroWatch, 2014
- Web server FOSS market lead (31.2%) — W3 Techs, 2014

Debian — the Project

Group of people united by a common goal:

Create the best, Free operating system.

Debian Social Contract (excerpt) (1997)

1. 100% Free Software → Debian Free Software Guidelines (DFSG)
2. give back
3. don't hide problems

- ≈ 1'000 official members world-wide
- ≈ 5'000 contributors
- volunteers, no employees

Compliance in Debian

Non-issues

common compliance issues that do *not* apply here:

- ~~*“release everything but your secret sauce”*~~
  - Free Software commitment: we *want* to release everything
- ~~copyright assignment / contributor license agreement~~

  exceptions:
  - responsibility waiving (e.g., *post mortem* license upgrades)
  - delegate license enforcement to trusted 3rd parties

Actual compliance issues

- keep Debian (main) 100% DFSG-free (mission)
- keep Debian mirrors content re-distributable (legal)
  - e.g., avoid license incompatibilities

debian/copyright

- human readable file that collects all copyright & license notices for any given (source) package[1]

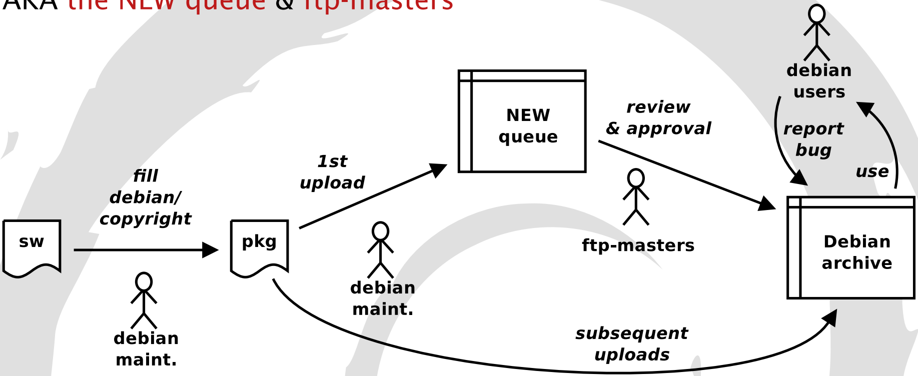
- developers: must fill it in, reviewing upstream notices
- users: for any given (binary) package PKG, will find it under `/usr/share/doc/PKG/copyright`
- popular licenses' full texts are collected under `/usr/share/common-licenses/` and referenced from debian/copyright

- incorrect debian/copyright → release critical bug

[1] www.debian.org/doc/debian-policy/ch-docs.html#s-copyrightfile

Review process

AKA the NEW queue & ftp-masters



- 2-tier review process
- main purpose: check DFSG free-ness

Machine-readable (M-R) debian/copyright

2007 early versions
2012 version 1.0

http://www.debian.org/doc/packaging-manuals/copyright-format/1.0/

```
Format: http://www.debian.org/doc/packaging-manuals/copyright-format/1.0/
Upstream-Name: X Solitaire
Source: ftp://ftp.example.com/pub/games

Files: *
Copyright: Copyright 1998 John Doe <jdoe@example.com>
License: GPL-2+
 This program is free software; you can redistribute it and/or modify it under the terms of the
 GNU General Public License as published by the Free Software Foundation; [snip]
 .
 On Debian systems, the full text of the GNU General Public License version 2 can be found
 in the file ‘/usr/share/common-licenses/GPL-2’.

Files: complex-1/*
Copyright: Copyright 1998 Jane Smith <jsmith@example.net>
License: GPL-2+ with OpenSSL exception
 [LICENSE TEXT]

Files: complex-2/*
Copyright: Copyright 1998 Jane Smith <jsmith@example.net>
License: GPL-2+ or Artistic-2.0, and BSD
 [LICENSE TEXT]
```

M-R debian/copyright — coverage

Potential: huge corpus of thrice reviewed copyright/license notices for popular Free Software projects, pluggable into your compliance processes.

Archive coverage of machine-readable debian/copyright files:[2]

date	release	source packages	archive coverage
Feb 2011	Squeeze	≈ 2’800	19%
May 2013	Wheezy	≈ 7’400	42%
Jan 2014	*unstable*	≈ 9’700	46%
Oct 2015	*unstable*	≈ 15’800	65%

[2] note: *all* (100%) Debian packages have a debian/copyright file, but not all of them are written in the machine-readable format (yet)

Outline

1. Debian and compliance
2. Debsources
3. Debsources for compliance

Debsources in a nutshell

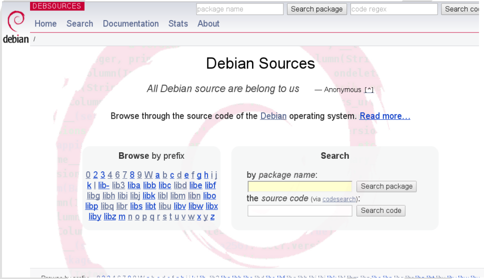
1. an infrastructure to publish Debian source code on the Web
2. a notable instance indexing *all* Debian source code to date: `http://sources.debian.net`

For developers:

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

For data miners:

- Debian evolution over time
- 20+ years of FOSS history
- live change monitoring



Debian Sources

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

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

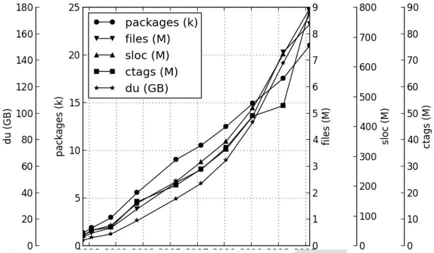
Browse by prefix

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 — data flow

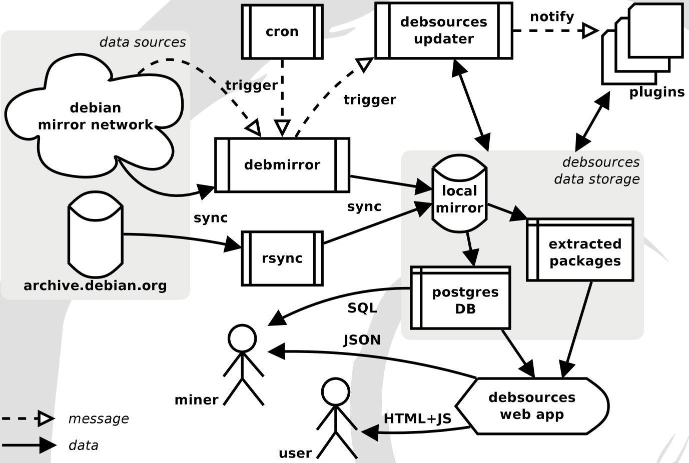


Figure: Debsources architecture

plugins

- disk usage
- sloccount
- ctags
- checksums (SHA256)

Debsources — coverage

Covered releases:

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

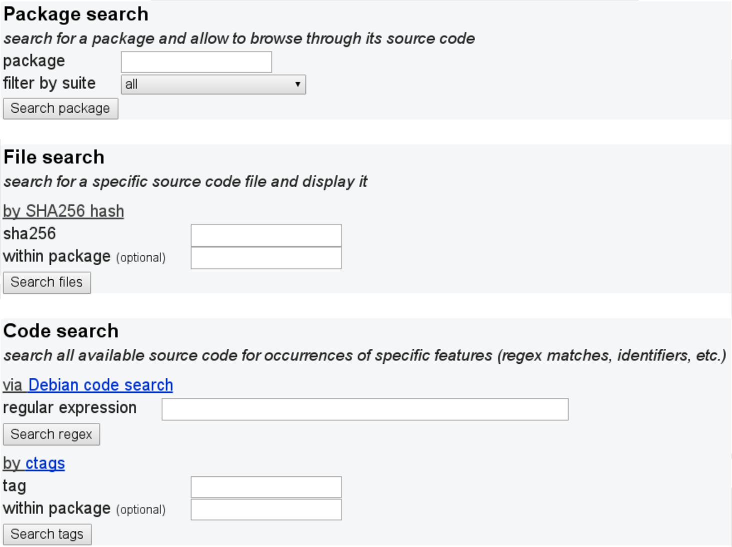
Update frequency: 4 times a day (at each Debian archive change)

Overall content: (Oct 2015)

- 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 — search



Package search

search for a package and allow to browse through its source code

package

filter by suite: all

File search

search for a specific source code file and display it

by SHA256 hash

sha256

within package (optional)

Code search

search all available source code for occurrences of specific features (regex matches, identifiers, etc.)

via Debian code search

regular expression

by ctags

tag

within package (optional)

Figure: http://sources.debian.net/advancedsearch/

Outline

1. Debian and compliance

2. Debsources

3. Debsources for compliance

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 files available 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)
- forthcoming: AND-ed ctags searches, file name searches

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/1.05.000-1/
http://sourcesdev.debian.net/copyright/spdx/gnubg/1.05.000-1/
```

```
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: alpha, dev. preview

Future work

- toolchain integration
  - source code scanner with Debsources as backend. . .
  - . . . and SPDX output
- SPDX feedback
  - we have feedback about SPDX adoption in Debsources
  - we want feedback from SPDX users about our export
- multiple license oracles
  - Debsources data model supports multiple "license oracles"
  - we have processed all stable releases with FOSSology and Ninka[3]
  - next: customizable SPDX exports based on oracle outputs
    "what's Ninka's take on Debian package webkit/1.8.1-3.4?"

[3] joint work with Daniel M. Germán

An industry-neutral compliance DB ?

- tooling is only as useful as its data quality
- debian/copyright quality
  - has served Debian well for 20+ years
  - . . . but you might find bugs!
  - please report them; it will help us helping you
- Debian is rather unique in this space
  - prominent FOSS vendor, non-profit, driven by FOSS ideals
  - ⇒ a natural hosting place for an industry-neutral compliance DB
  - The Debian community is doing their part.
    With your contributions Debian might become a fundamental compliance hub. For everybody.

- As a proxy for Debian, Debsources is a useful resource to add to your mix of compliance data sources
- As a non-profit actor, Debian is in a sweet spot to federate compliance information for popular FOSS projects
- Let's discuss *your* use cases and how we can address them!

Thanks!

Questions?

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

about the slides:
available at https://upsilon.cc/~zack/talks/2015/20151008-lf-debsources.pdf