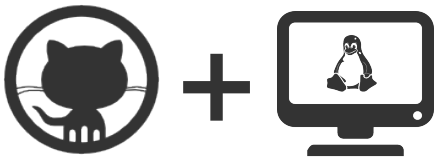


=> 1 million SPDX

Large-scale license transparency using open data, open standards and F/OSS

http://triplecheck.net

http://searchcode.com

Speaker

Nuno Brito

- Free/open source contributor since 2005
- Last 12 months wrote 100k F/OSS lines of code
- SPDX contributor, co-founder of TripleCheck

Around the web

http://nunobrito.eu

Transparency

Take some source code as example

Who developed the code?
Which licenses are applicable?
Was the code copied from somewhere else?

Size

A problem of scale

Open licenses? > 300 types to choose
> 5 million F/OSS projects
> 100 million source code files

Practice

Applying licenses

- Burden on developer *(do correctly, do enough)*
- Expressed differently *(difficult to understand)*
- Scaling obstacles *(scarce automation)*

Transparency?

What do?

Ideally, we'd have tooling that is..

a) Reachable
b) Cooperative
c) Free

Choose two. (sad reality)

Choose three

Choose building blocks based on:

a) Open standards
b) Open data
c) Reachable tools

Learn, write, improve.

Share.

Standards

SPDX: Open standard for software licensing

- Standardizes license description
- Defines Id for license terms
- http://spdx.org

Pro: Good docs, straightforward, getting better
Cons: Slow adoption, scarce tooling

Open data

GitHub: Targeting open data repositories

- API suited for intensive access
- Social coding
- Largest open source code collection

Pro: Reachable, diverse
Cons: Repositories processed one-by-one

Tooling

Custom-built tools for software licenses

- Large-scale repository data-mining
- Find applicable licenses inside content
- Share millions of SPDX documents

Pro: Learn by doing, modularized, single language
Cons: Built from scratch, needs consolidation

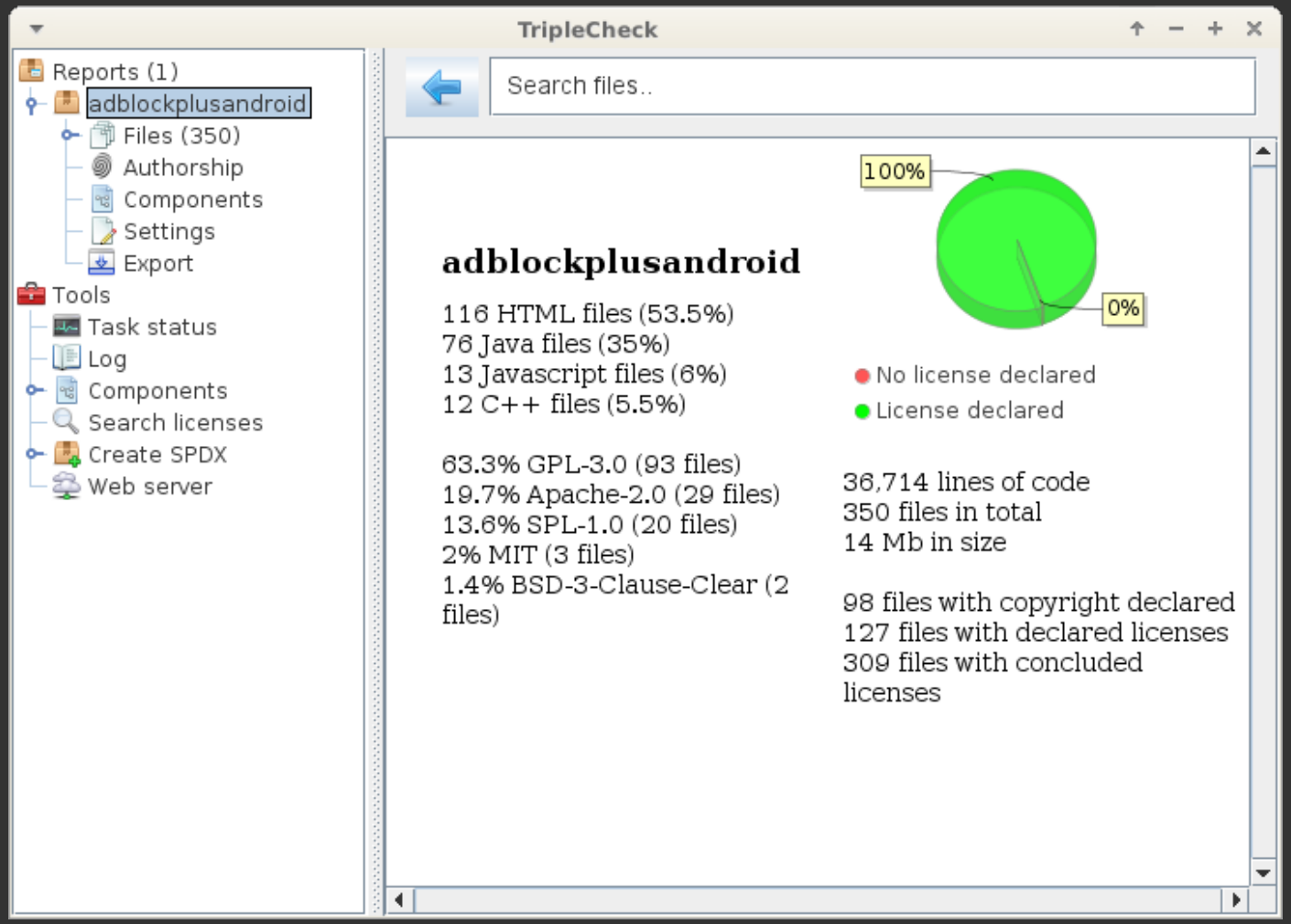
Step 1

Desktop tool/engine to discover licenses

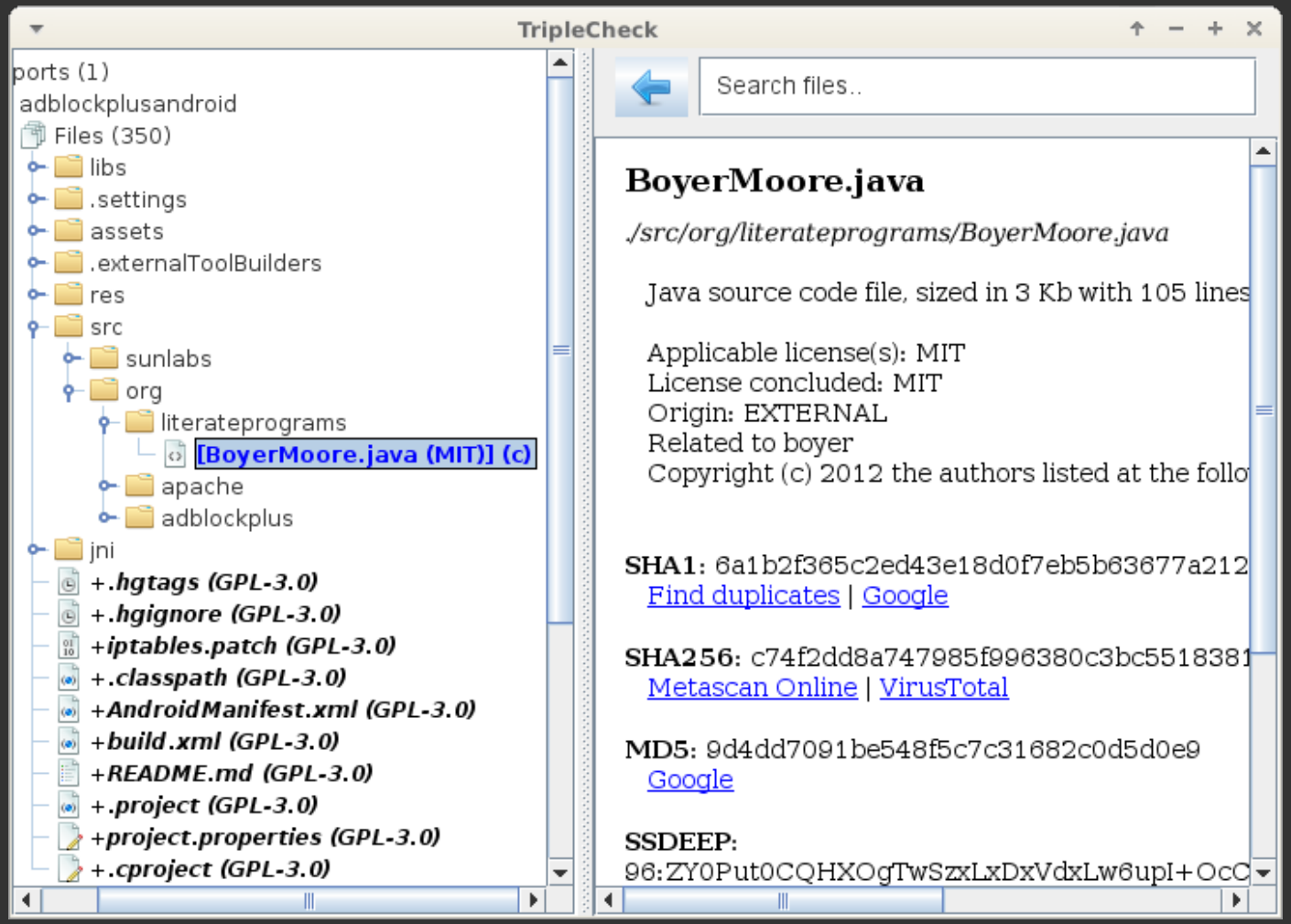
- SPDX format as storage medium
- Identify copyright and 18 license types
- Java, released in Feb 2014. EUPL

http://spdx.org/tools/community/triplecheck-reporter

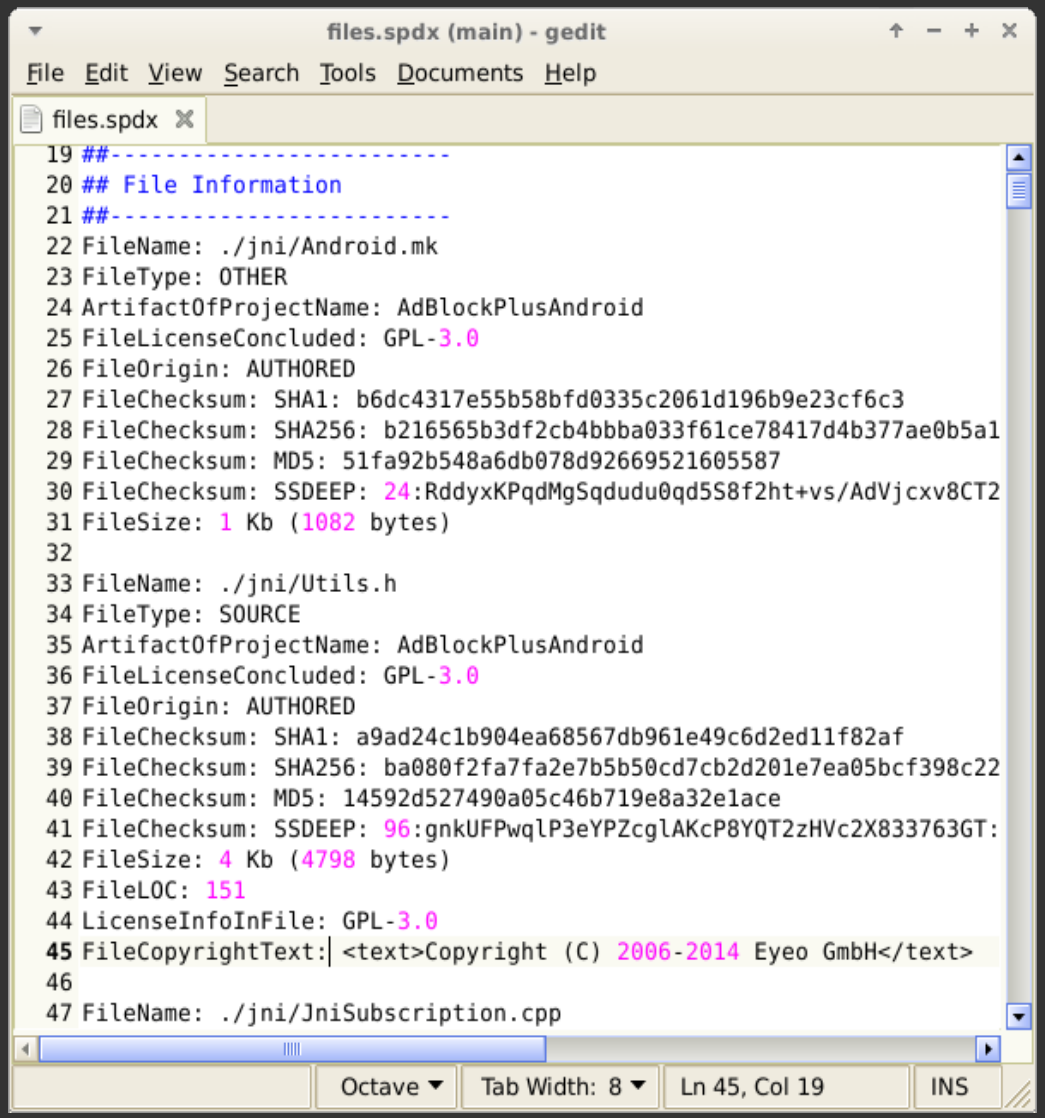
Desktop


TripleCheck
Reports (1)
adblockplusandroid
Files (350)
Authorship
Components
Settings
Export
Tools
Task status
Log
Components
Search licenses
Create SPDX
Web server
Search files..
adblockplusandroid
116 HTML files (53.5%)
76 Java files (35%)
13 Javascript files (6%)
12 C++ files (5.5%)
63.3% GPL-3.0 (93 files)
19.7% Apache-2.0 (29 files)
13.6% SPL-1.0 (20 files)
2% MIT (3 files)
1.4% BSD-3-Clause-Clear (2 files)
100%
0%
No license declared
License declared
36,714 lines of code
350 files in total
14 Mb in size
98 files with copyright declared
127 files with declared licenses
309 files with concluded licenses

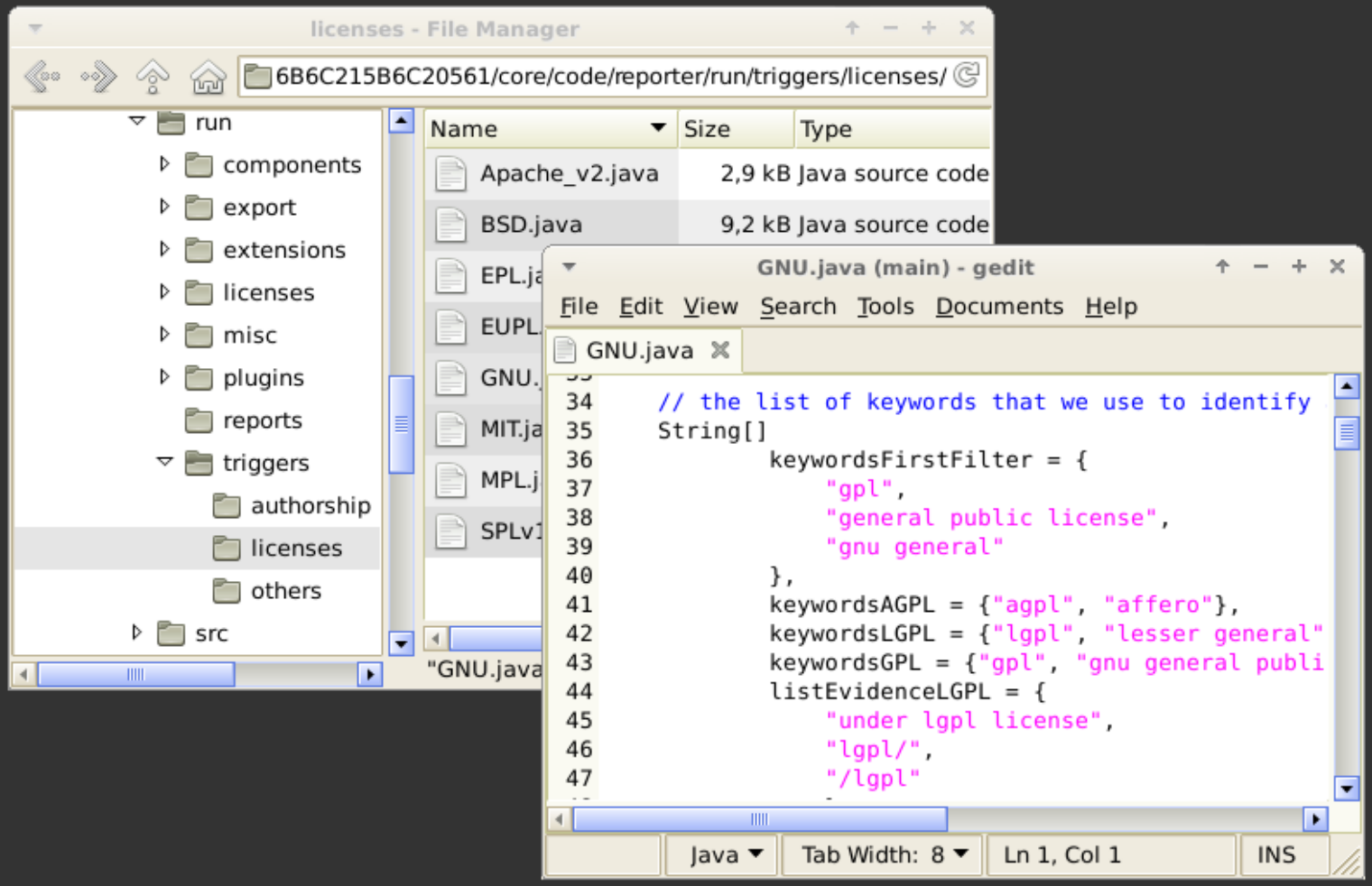
File detail


TripleCheck
ports (1)
adblockplusandroid
Files (350)
libs
.settings
assets
.externalToolBuilders
res
src
sunlabs
org
literateprograms
[BoyerMoore.java (MIT)] (c)
apache
adblockplus
jni
+.hgtags (GPL-3.0)
+.hgignore (GPL-3.0)
+iptables.patch (GPL-3.0)
+.classpath (GPL-3.0)
+AndroidManifest.xml (GPL-3.0)
+build.xml (GPL-3.0)
+README.md (GPL-3.0)
+.project (GPL-3.0)
+project.properties (GPL-3.0)
+.cproject (GPL-3.0)
Search files..
BoyerMoore.java
./src/org/literateprograms/BoyerMoore.java
Java source code file, sized in 3 Kb with 105 lines
Applicable license(s): MIT
License concluded: MIT
Origin: EXTERNAL
Related to boyer
Copyright (c) 2012 the authors listed at the follo
SHA1: 6a1b2f365c2ed43e18d0f7eb5b63677a212
Find duplicates | Google
SHA256: c74f2dd8a747985f996380c3bc5518381
Metascan Online | VirusTotal
MD5: 9d4dd7091be548f5c7c31682c0d5d0e9
Google
SSDEEP:
96:ZY0Put0CQHXOgTwSzxLxDxVdxLw6upI+OcC

SPDX file


files.spdx (main) - gedit
File Edit View Search Tools Documents Help
files.spdx
19 ##------------------------
20 ## File Information
21 ##------------------------
22 FileName: ./jni/Android.mk
23 FileType: OTHER
24 ArtifactOfProjectName: AdBlockPlusAndroid
25 FileLicenseConcluded: GPL-3.0
26 FileOrigin: AUTHORED
27 FileChecksum: SHA1: b6dc4317e55b58bfd0335c2061d196b9e23cf6c3
28 FileChecksum: SHA256: b216565b3df2cb4bbba033f61ce78417d4b377ae0b5a1
29 FileChecksum: MD5: 51fa92b548a6db078d92669521605587
30 FileChecksum: SSDEEP: 24:RddyxKPqdMgSqdudu0qd5S8f2ht+vs/AdVjcxv8CT2
31 FileSize: 1 Kb (1082 bytes)
32
33 FileName: ./jni/Utils.h
34 FileType: SOURCE
35 ArtifactOfProjectName: AdBlockPlusAndroid
36 FileLicenseConcluded: GPL-3.0
37 FileOrigin: AUTHORED
38 FileChecksum: SHA1: a9ad24c1b904ea68567db961e49c6d2ed11f82af
39 FileChecksum: SHA256: ba080f2fa7fa2e7b5b50cd7cb2d201e7ea05bcf398c22
40 FileChecksum: MD5: 14592d527490a05c46b719e8a32e1ace
41 FileChecksum: SSDEEP: 96:gnkUFPwqlP3eYPZcglAKcP8YQT2zHVc2X833763GT:
42 FileSize: 4 Kb (4798 bytes)
43 FileLOC: 151
44 LicenseInfoInFile: GPL-3.0
45 FileCopyrightText: <text>Copyright (C) 2006-2014 Eyeo GmbH</text>
46
47 FileName: ./jni/JniSubscription.cpp
Octave
Tab Width: 8
Ln 45, Col 19
INS

Customize


licenses - File Manager
6B6C215B6C20561/core/code/reporter/run/triggers/licenses/
run
components
export
extensions
licenses
misc
plugins
reports
triggers
authorship
licenses
others
src
Name
Size
Type
Apache_v2.java
2,9 kB Java source code
BSD.java
9,2 kB Java source code
GNU.java (main) - gedit
File Edit View Search Tools Documents Help
GNU.java
// the list of keywords that we use to identify
String[]
keywordsFirstFilter = {
"gpl",
"general public license",
"gnu general"
},
keywordsAGPL = {"agpl", "affero"},
keywordsLGPL = {"lgpl", "lesser general"
keywordsGPL = {"gpl", "gnu general publi
listEvidenceLGPL = {
"under lgpl license",
"lgpl/",
"/lgpl"
Java
Tab Width: 8
Ln 1, Col 1
INS

Details

Underneath the hood

- 147 file extensions, 18 license types
- LOC, hashes *(SHA1, MD5, SHA256, SSDEEP)*
- Command line supported *(Jenkins, cron)*
- Fast, 40k files/minute *(Pentium IV)*

Step 2

Discovering repositories with gitFinder

Create a list of projects online to use as components.
Get basic licensing information from each project.

- Write text file with each github user (~7 million)
- For each user, find repositories not forked (~10M)
- Split each repository according to language (197)
- For each list of language/reps, download code

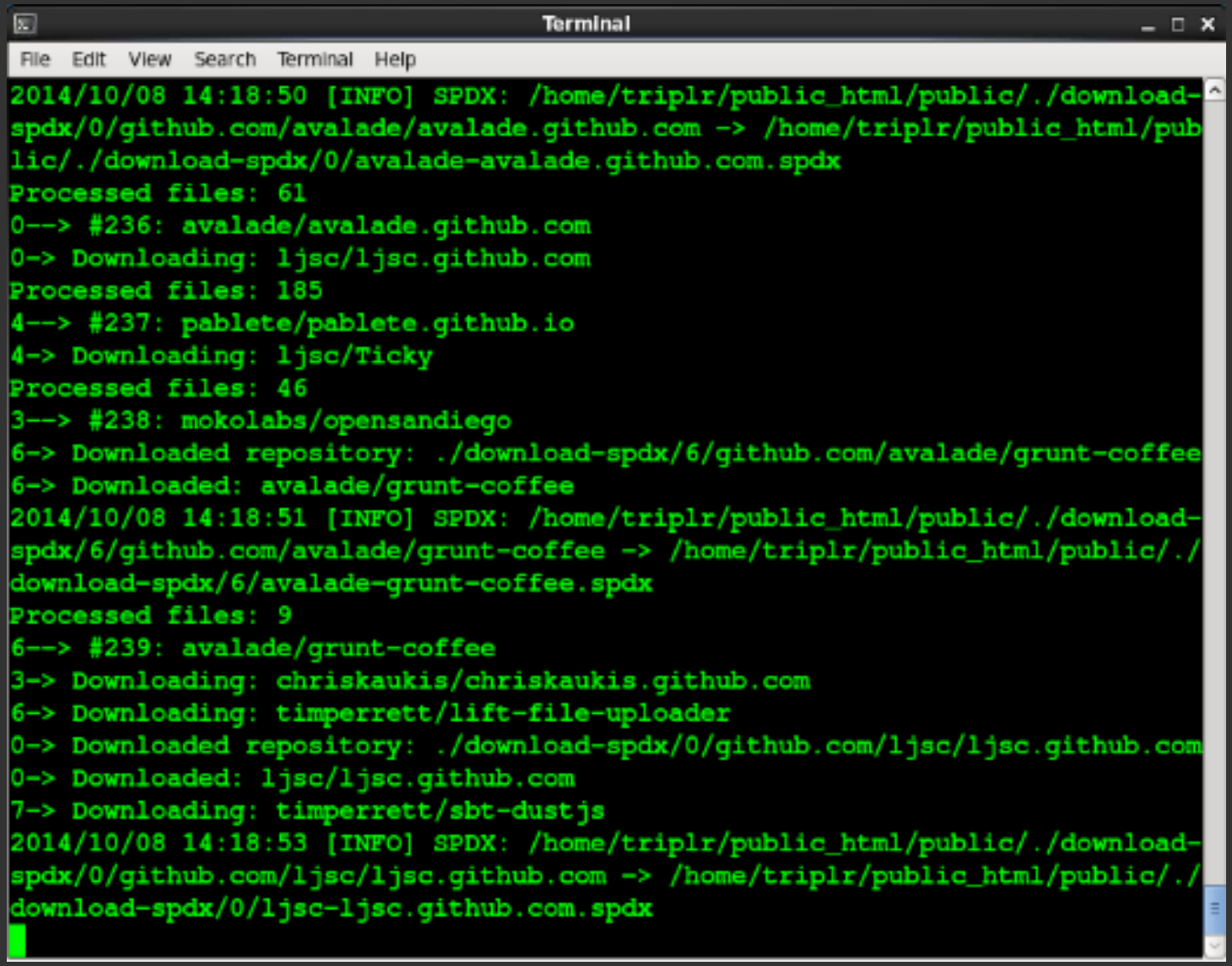
Performance

~70k repositories/day

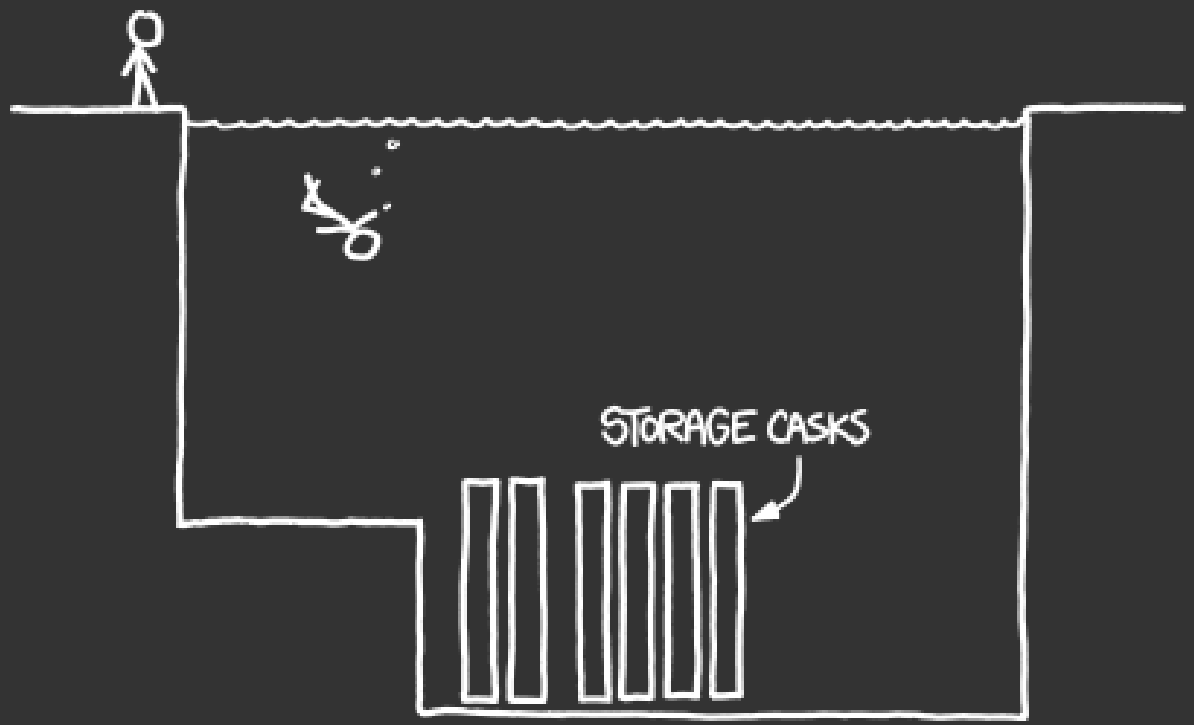
- Single machine (i7, 8Gb RAM, CentOS)
- 9 parallel threads
- Resume/recover supported
- Released in Jun. 2014

https://github.com/triplecheck/gitfinder

Output


Terminal
File Edit View Search Terminal Help
2014/10/08 14:18:50 [INFO] SPDX: /home/triplr/public_html/public/./download-spdx/0/github.com/avalade/avalade.github.com -> /home/triplr/public_html/public/./download-spdx/0/avalade-avalade.github.com.spdx
Processed files: 61
0--> #236: avalade/avalade.github.com
0-> Downloading: ljsc/ljsc.github.com
Processed files: 185
4--> #237: pablete/pablete.github.io
4-> Downloading: ljsc/Ticky
Processed files: 46
3--> #238: mokolabs/opensandiego
6-> Downloaded repository: ./download-spdx/6/github.com/avalade/grunt-coffee
6-> Downloaded: avalade/grunt-coffee
2014/10/08 14:18:51 [INFO] SPDX: /home/triplr/public_html/public/./download-spdx/6/github.com/avalade/grunt-coffee -> /home/triplr/public_html/public/./download-spdx/6/avalade-grunt-coffee.spdx
Processed files: 9
6--> #239: avalade/grunt-coffee
3-> Downloading: chriskaukis/chriskaukis.github.com
6-> Downloading: timperrett/lift-file-uploader
0-> Downloaded repository: ./download-spdx/0/github.com/ljsc/ljsc.github.com
0-> Downloaded: ljsc/ljsc.github.com
7-> Downloading: timperrett/sbt-dustjs
2014/10/08 14:18:53 [INFO] SPDX: /home/triplr/public_html/public/./download-spdx/0/github.com/ljsc/ljsc.github.com -> /home/triplr/public_html/public/./download-spdx/0/ljsc-ljsc.github.com.spdx

Storage?



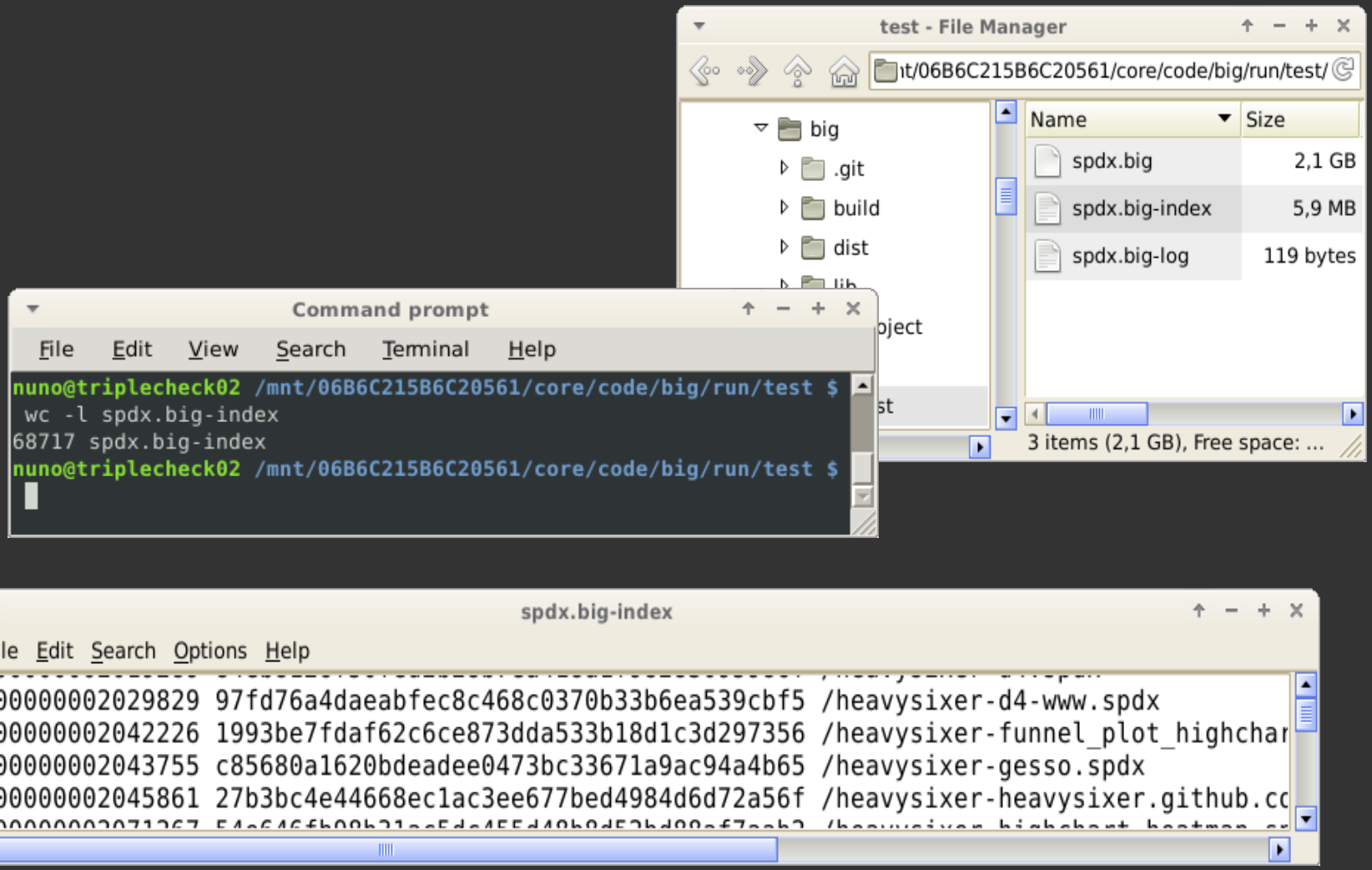
Storage

BigZip, +100 million files on a single download

- Flat-file, zip compression *(per entry)*
- Fast, simple, portable. Indexed search

https://github.com/triplecheck/big

How it looks


test - File Manager
/06B6C215B6C20561/core/code/big/run/test/
big
.git
build
dist
Name
Size
spdx.big
2,1 GB
spdx.big-index
5,9 MB
spdx.big-log
119 bytes
3 items (2,1 GB), Free space: ...
Command prompt
File Edit View Search Terminal Help
nuno@triplecheck02 /mnt/06B6C215B6C20561/core/code/big/run/test $
wc -l spdx.big-index
68717 spdx.big-index
nuno@triplecheck02 /mnt/06B6C215B6C20561/core/code/big/run/test $
spdx.big-index
File Edit Search Options Help
000000002029829 97fd76a4daeabfec8c468c0370b33b6ea539cbf5 /heavysixer-d4-www.spdx
000000002042226 1993be7fdaf62c6ce873dda533b18d1c3d297356 /heavysixer-funnel_plot_highchar
000000002043755 c85680a1620bdeadee0473bc33671a9ac94a4b65 /heavysixer-gesso.spdx
000000002045861 27b3bc4e44668ec1ac3ee677bed4984d6d72a56f /heavysixer-heavysixer.github.cc

Step 3

SPDX search engine

- One-click SPDX creation from open data
- Visualize license and copyright data
- Visit at http://searchcode.com/spdx

Example

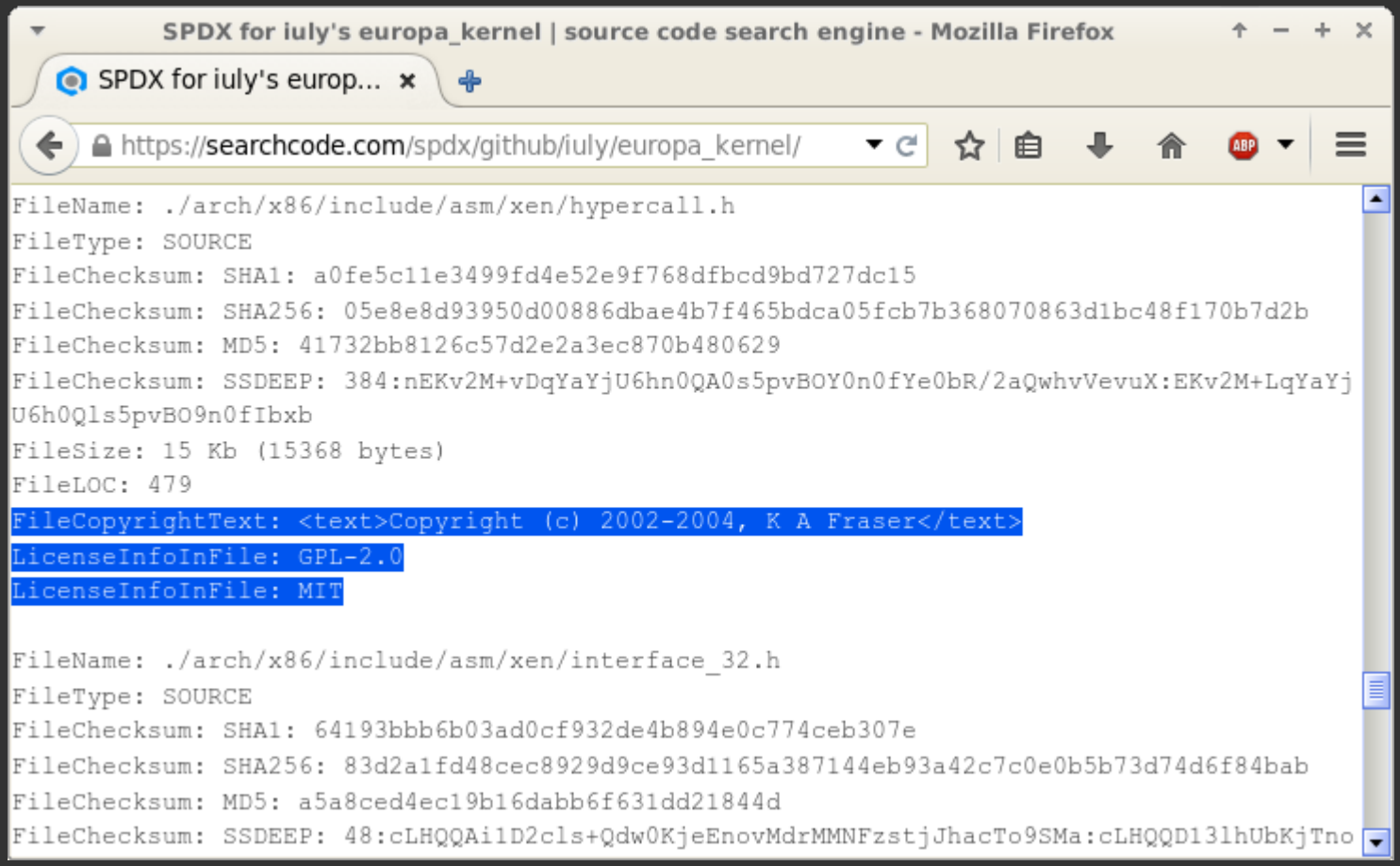
Using the original URL..

- https://**git**hub.com/iuly/europa_kernel/

  =>

- https://**spdx**hub.com/iuly/europa_kernel/

Example


SPDX for iuly's europa_kernel | source code search engine - Mozilla Firefox
SPDX for iuly's europ...
https://searchcode.com/spdx/github/iuly/europa_kernel/
FileName: ./arch/x86/include/asm/xen/hypercall.h
FileType: SOURCE
FileChecksum: SHA1: a0fe5c11e3499fd4e52e9f768dfbcd9bd727dc15
FileChecksum: SHA256: 05e8e8d93950d00886dbae4b7f465bdca05fcb7b368070863d1bc48f170b7d2b
FileChecksum: MD5: 41732bb8126c57d2e2a3ec870b480629
FileChecksum: SSDEEP: 384:nEKv2M+vDqYaYjU6hn0QA0s5pvBOY0n0fYe0bR/2aQwhvVevuX:EKv2M+LqYaYj
U6h0Qls5pvBO9n0fIbxb
FileSize: 15 Kb (15368 bytes)
FileLOC: 479
FileCopyrightText: <text>Copyright (c) 2002-2004, K A Fraser</text>
LicenseInfoInFile: GPL-2.0
LicenseInfoInFile: MIT
FileName: ./arch/x86/include/asm/xen/interface_32.h
FileType: SOURCE
FileChecksum: SHA1: 64193bbb6b03ad0cf932de4b894e0c774ceb307e
FileChecksum: SHA256: 83d2a1fd48cec8929d9ce93d1165a387144eb93a42c7c0e0b5b73d74d6f84bab
FileChecksum: MD5: a5a8ced4ec19b16dabb6f631dd21844d
FileChecksum: SSDEEP: 48:cLHQQAi1D2cls+Qdw0KjeEnovMdrMMNFzstjJhacTo9SMa:cLHQQD13lhUbKjTno

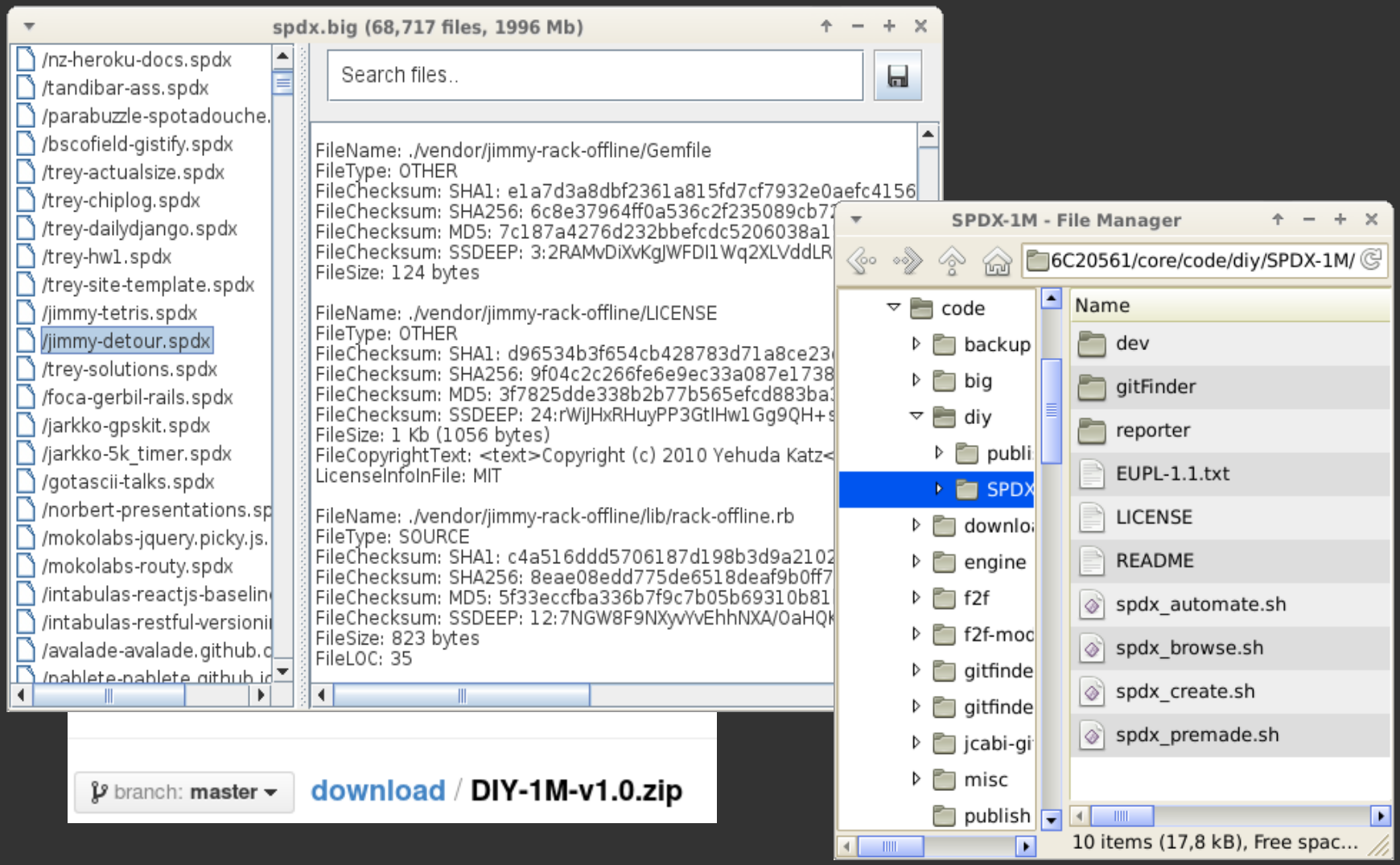
SPDX-1M

“Do It Yourself” kit. Generate 1 million SPDX

- https://github.com/triplecheck/diy
- 1.2 million open source projects
- “Arduino” for s/w licenses detection

9Gb worth of SPDX? Grab:
http://triplecheck.net/public/storage/spdx.big

Screenshots



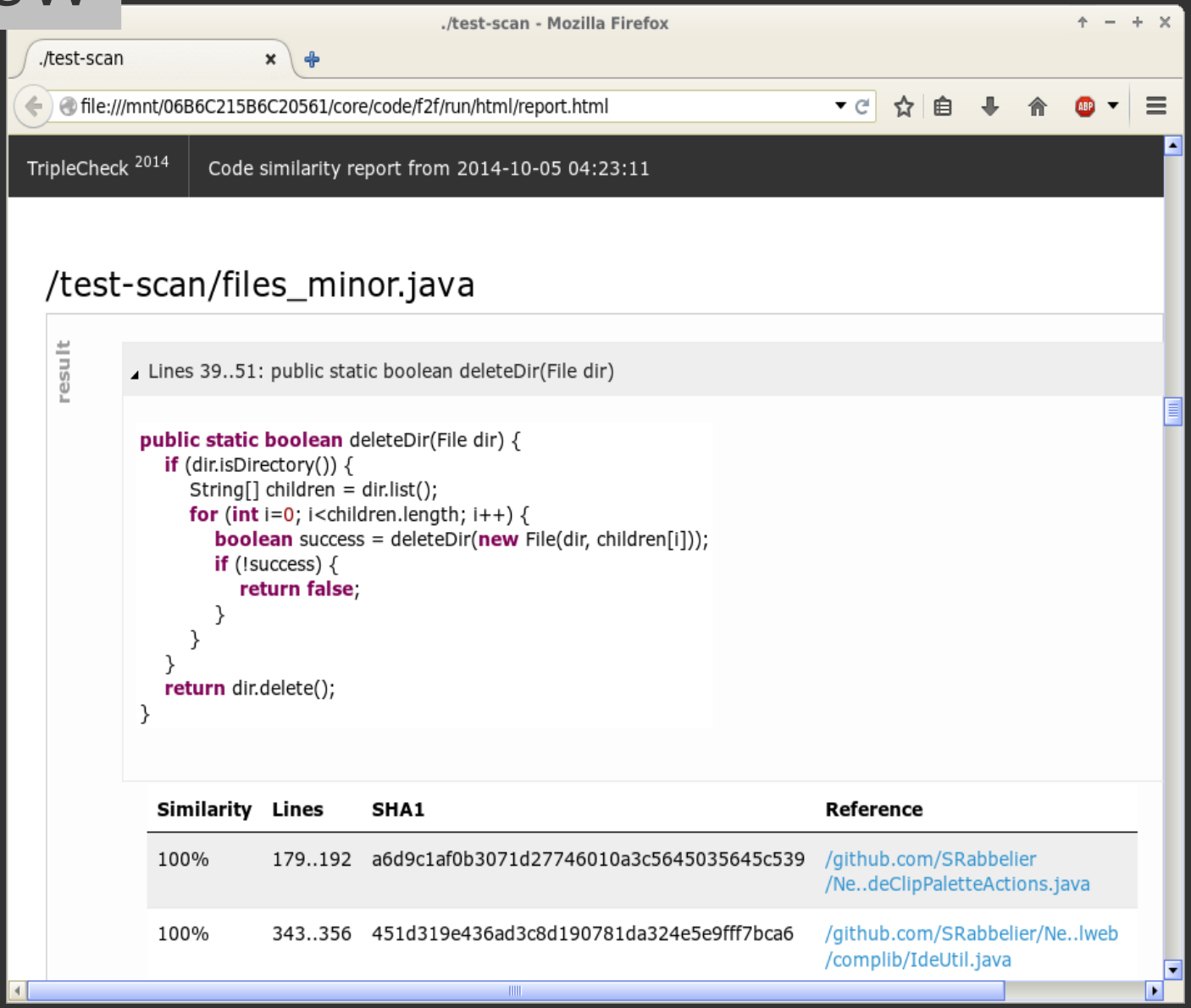
Next step?

F2F – pinpointing non-original code

- Decompose code into blocks
- Tokenize/anonymize data
- Find code matches across knowledge base

ETA in Dec. 2014
https://github.com/triplecheck/f2f

Preview


./test-scan - Mozilla Firefox
./test-scan
file:///mnt/06B6C215B6C20561/core/code/f2f/run/html/report.html
TripleCheck 2014
Code similarity report from 2014-10-05 04:23:11
/test-scan/files_minor.java
result
Lines 39..51: public static boolean deleteDir(File dir)
public static boolean deleteDir(File dir) {
if (dir.isDirectory()) {
String[] children = dir.list();
for (int i=0; i<children.length; i++) {
boolean success = deleteDir(new File(dir, children[i]));
if (!success) {
return false;
}
}
}
return dir.delete();
}
Similarity
Lines
SHA1
Reference
100%
179..192
a6d9c1af0b3071d27746010a3c5645035645c539
/github.com/SRabbelier /Ne..deClipPaletteActions.java
100%
343..356
451d319e436ad3c8d190781da324e5e9fff7bca6
/github.com/SRabbelier/Ne..lweb /complib/IdeUtil.java

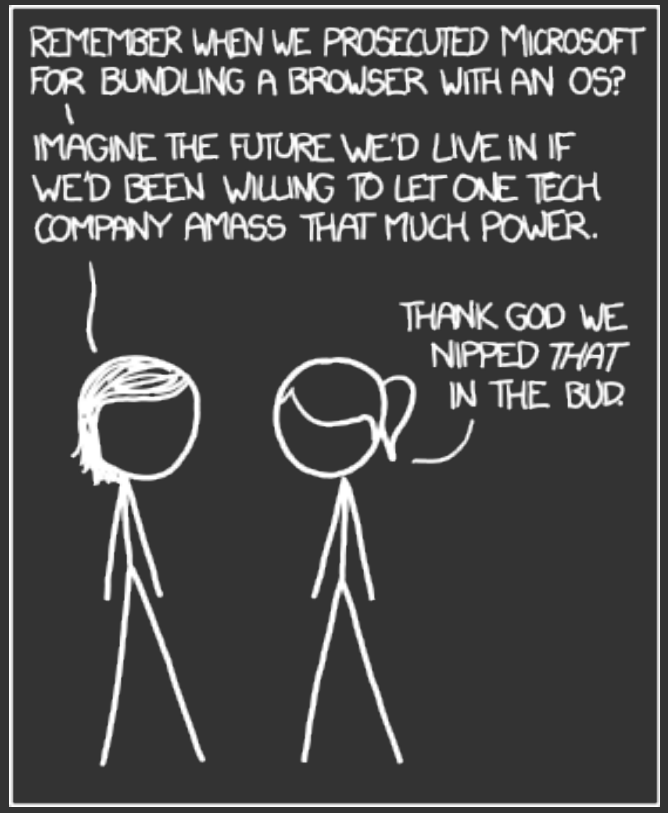
Conclusion

What is now available for everyone

- Desktop tooling / detection engine
- Extraction of open data in scale
- Search engine for SPDX

Questions?

http://spdx.org
http://searchcode.com/spdx
http://github.com/triplecheck



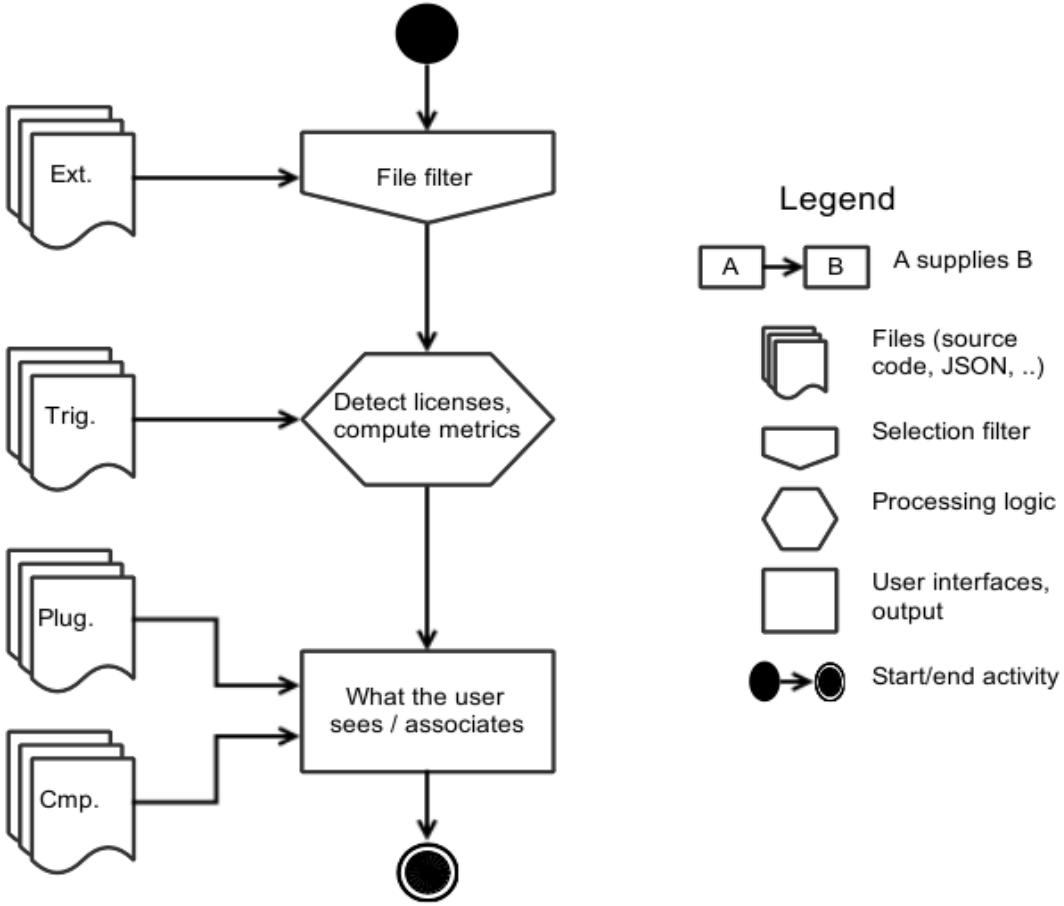
http://xkcd.com/1118/

Interesting stuff?

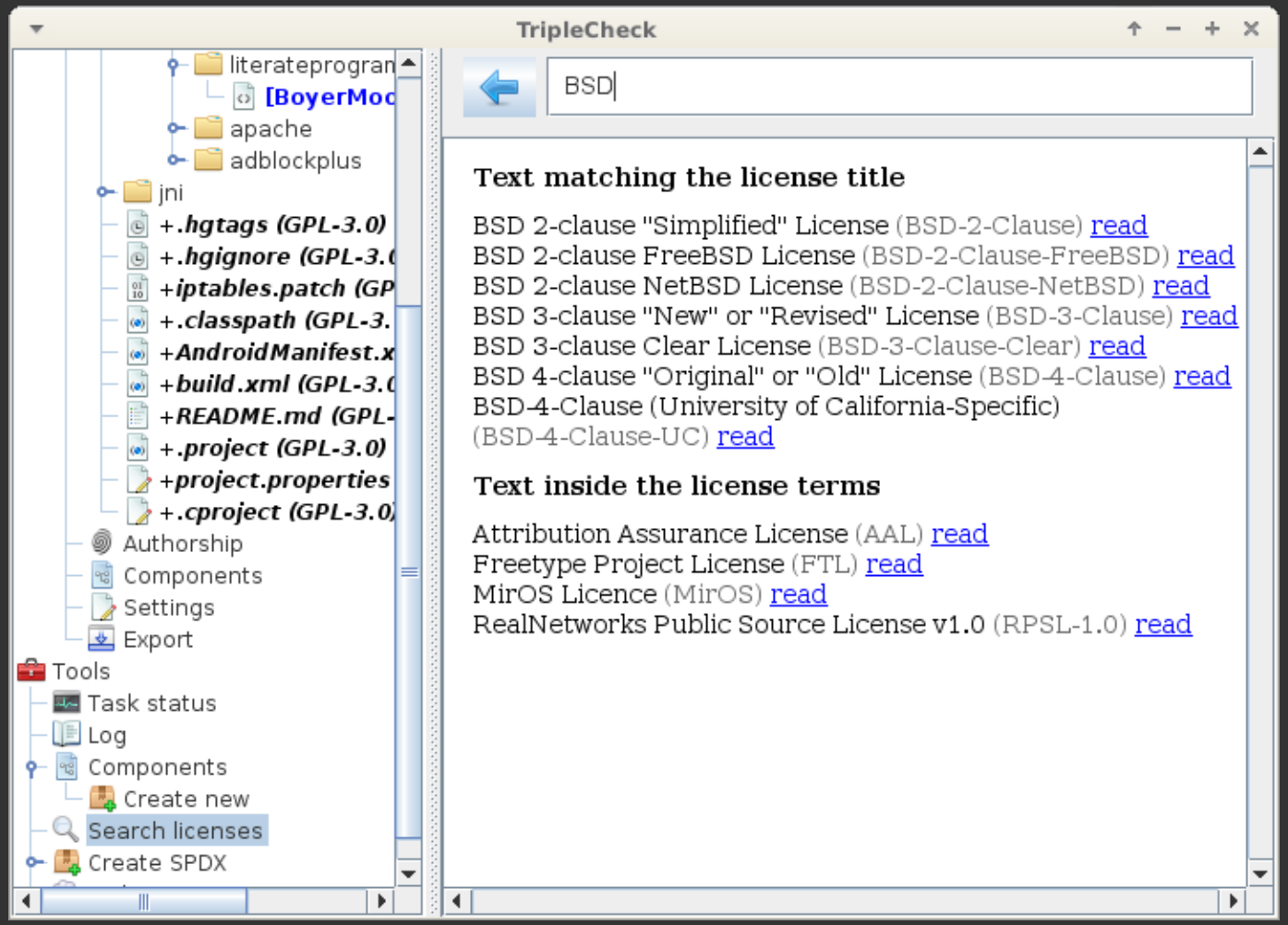
Let us know: @nn81 @boyte #linuxcon

Backup slides

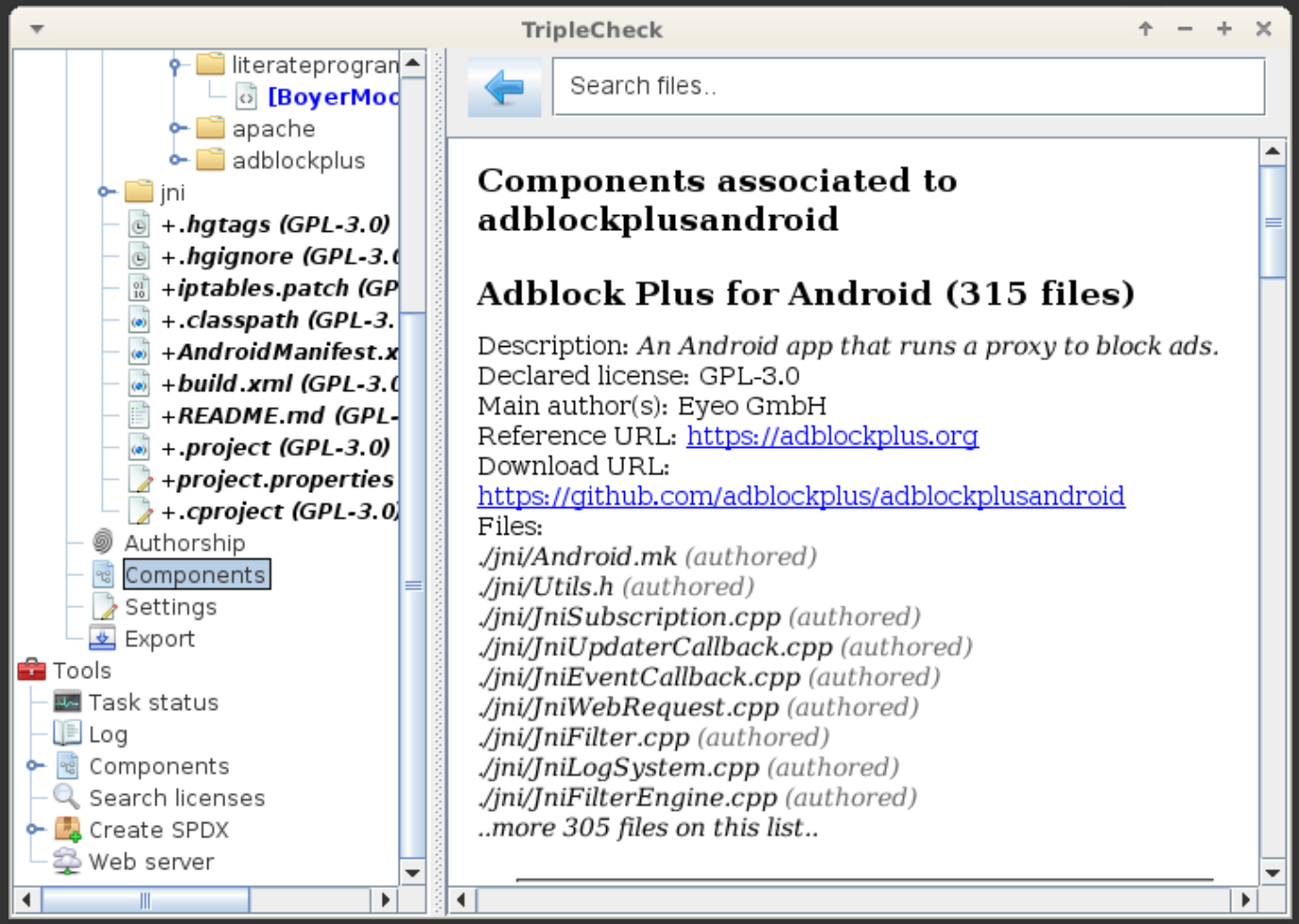
Engine


Ext.
File filter
Trig.
Detect licenses, compute metrics
Plug.
Cmp.
What the user sees / associates
Legend
A
B
A supplies B
Files (source code, JSON, ..)
Selection filter
Processing logic
User interfaces, output
Start/end activity

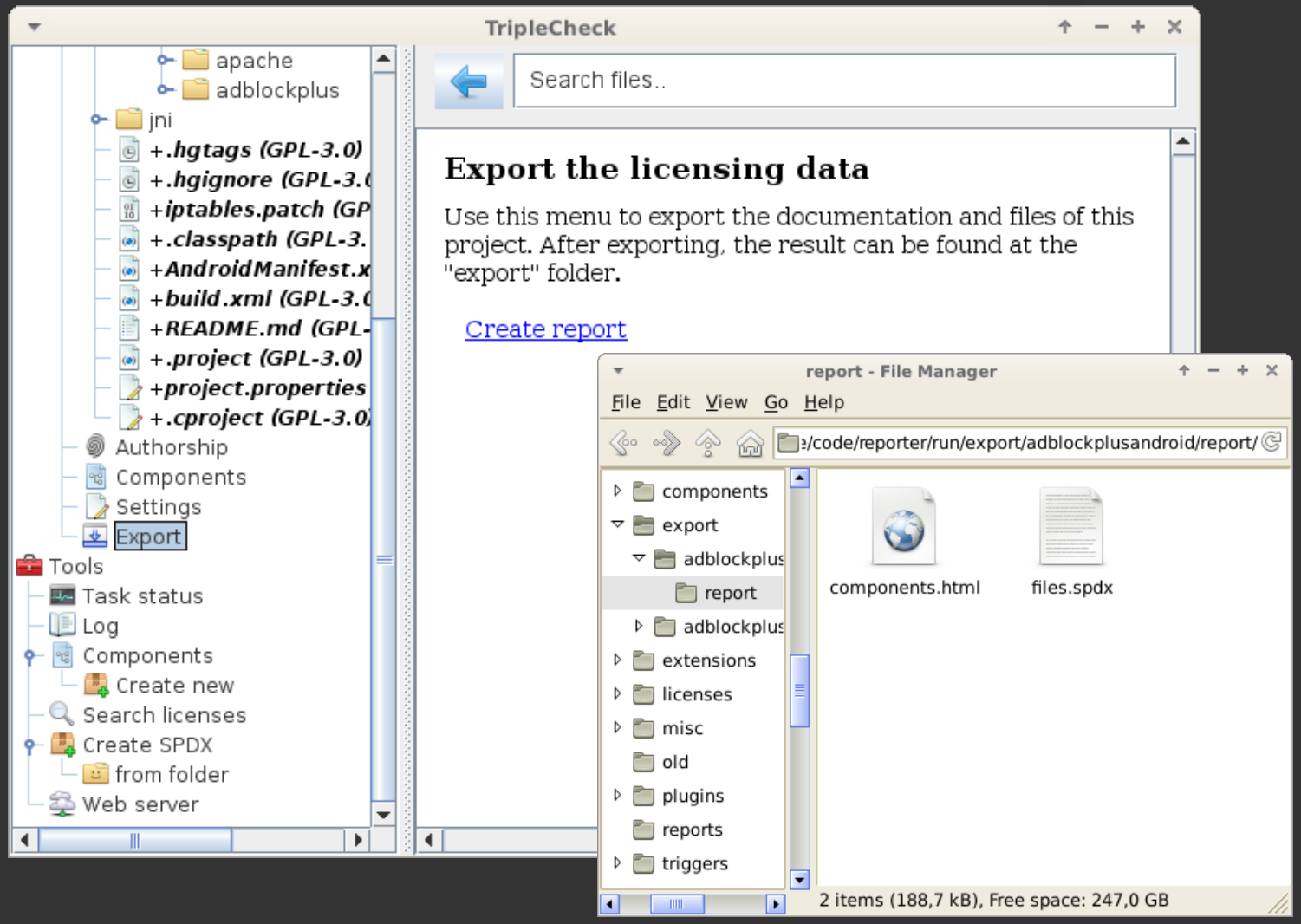
License DB


TripleCheck
literateprogran
[BoyerMoc
apache
adblockplus
jni
+.hgtags (GPL-3.0)
+.hgignore (GPL-3.
+iptables.patch (GP
+.classpath (GPL-3.
+AndroidManifest.x
+build.xml (GPL-3.
+README.md (GPL-
+.project (GPL-3.0)
+project.properties
+.cproject (GPL-3.0)
Authorship
Components
Settings
Export
Tools
Task status
Log
Components
Create new
Search licenses
Create SPDX
BSD
Text matching the license title
BSD 2-clause "Simplified" License (BSD-2-Clause) read
BSD 2-clause FreeBSD License (BSD-2-Clause-FreeBSD) read
BSD 2-clause NetBSD License (BSD-2-Clause-NetBSD) read
BSD 3-clause "New" or "Revised" License (BSD-3-Clause) read
BSD 3-clause Clear License (BSD-3-Clause-Clear) read
BSD 4-clause "Original" or "Old" License (BSD-4-Clause) read
BSD-4-Clause (University of California-Specific)
(BSD-4-Clause-UC) read
Text inside the license terms
Attribution Assurance License (AAL) read
Freetype Project License (FTL) read
MirOS Licence (MirOS) read
RealNetworks Public Source License v1.0 (RPSL-1.0) read

Components


TripleCheck
literateprogran
[BoyerMoo
apache
adblockplus
jni
+.hgtags (GPL-3.0)
+.hgignore (GPL-3.
+iptables.patch (GP
+.classpath (GPL-3.
+AndroidManifest.x
+build.xml (GPL-3.
+README.md (GPL-
+.project (GPL-3.0)
+project.properties
+.cproject (GPL-3.0)
Authorship
Components
Settings
Export
Tools
Task status
Log
Components
Search licenses
Create SPDX
Web server
Search files..
Components associated to adblockplusandroid
Adblock Plus for Android (315 files)
Description: An Android app that runs a proxy to block ads.
Declared license: GPL-3.0
Main author(s): Eyeo GmbH
Reference URL: https://adblockplus.org
Download URL:
https://github.com/adblockplus/adblockplusandroid
Files:
./jni/Android.mk (authored)
./jni/Utils.h (authored)
./jni/JniSubscription.cpp (authored)
./jni/JniUpdaterCallback.cpp (authored)
./jni/JniEventCallback.cpp (authored)
./jni/JniWebRequest.cpp (authored)
./jni/JniFilter.cpp (authored)
./jni/JniLogSystem.cpp (authored)
./jni/JniFilterEngine.cpp (authored)
..more 305 files on this list..

Exporting


TripleCheck
apache
adblockplus
jni
+.hgtags (GPL-3.0)
+.hgignore (GPL-3.0
+iptables.patch (GP
+.classpath (GPL-3.
+AndroidManifest.x
+build.xml (GPL-3.0
+README.md (GPL-
+.project (GPL-3.0)
+project.properties
+.cproject (GPL-3.0)
Authorship
Components
Settings
Export
Tools
Task status
Log
Components
Create new
Search licenses
Create SPDX
from folder
Web server
Search files..
Export the licensing data
Use this menu to export the documentation and files of this project. After exporting, the result can be found at the "export" folder.
Create report
report - File Manager
File Edit View Go Help
:/code/reporter/run/export/adblockplusandroid/report/
components
export
adblockplus
report
adblockplus
extensions
licenses
misc
old
plugins
reports
triggers
components.html
files.spdx
2 items (188,7 kB), Free space: 247,0 GB