FOSSology
SPDX in HD

Speaker: michael.c.jaeger@siemens.com

Siemens Corporate Technology

What is this about?

You know these examples

- It is about licensing of software packages
- Licensing is expressed in files
- You would like to find licensing statements …
- …and summarize this in SPDX

jakarta-oro-2.0.8.zip

```
<spdx:hasExtractedLicensingInfo>
<spdx:ExtractedLicensingInfo
rdf:about="http://spdx.org/licenses/Apache-
1.1">
<spdx:licenseId>Apache-
1.1</spdx:licenseId>
<spdx:extractedText>
<![CDATA[
The Apache Software License, Version 1.1
*
 * Copyright (c) 2000 The Apache Software
Foundation.  All rights
 * reserved.
 *
```
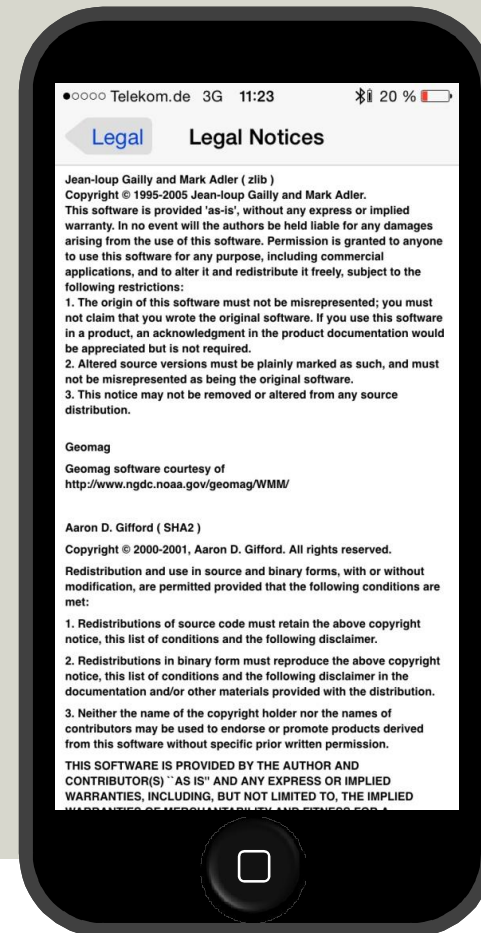
Why we are doing this?

You know these examples
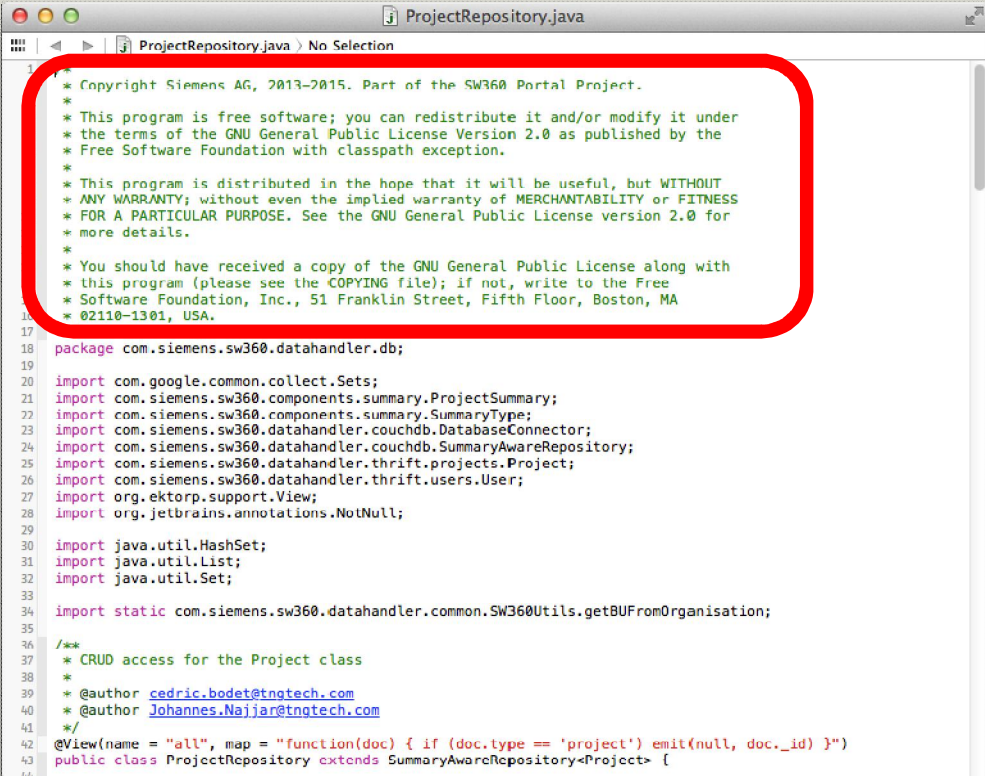
Distributing open source software requires to

- Provide licenses of involved software
- Provide copyright statements of involved authors
- Provide disclaimers
- … and much more

It is about finding licenses

Finding Licenses

- License texts
- References to licenses
- Written texts explaining licensing
- License relevant statements

An Example – What do we find?

SIEMENS

The Example

- (Apache projects are known for homogenous licensing)
- Such a project declares its licensing on its Website
- However, Open Source implies using other open source
- As such, a projects can contain also parts from other open source projects

Using FOSSology with this Example

Open Source and Reuse

- It is natural that an OSS project reuses available https://github.com/fossology/fossology
- Likely OSS from other projects is found
- For example, FOSSology will find 25 other licensing relevant text occurrences in Apache thrift

Using FOSSology with another Example

Another Example: Linux Kernel

- Looking at the Linux kernel will result in thousands of files containing license relevant information
- Analyzing them yields about 30 distinctive licensing statements

How does FOSSology work?

```
Upload component
        ▼
Agents Scanning
        ▼
Review Results
        ▼
Generate Reporting
        ▼
Pass Report to Client
```

- Uploading source code archive (*.zip, *.tar.gz, etc)

- Agents scan for license relevant text
- Copyrights, ECC, your keywords to look for etc.

- Review scanner results for wrong license classification
- Review other scanner findings (copyrights, ECC)

- Result of the "clearing"
    - SPDX reporting
    - Generated notice or readme file
    - debian-copyright

What is FOSSology?

A Web server application for license and copyright compliance of software components.

FOSSology Project

https://www.fossology.org/

- Published first in 2008, GPL-2.0
- 2015: Linux Foundation collaboration project
- Web server based and command line interfaces
- Scanning agents searching for license and copyright relevant hits (and more …)
- A multi-user / multi-tenant Web UI for review organizing clearing job

FOSSology Development

https://www.github.com/fossology/fossology

- Standard Web application stack:
  - Linux, Apache 2, PostgreSQL, PHP,
- Web-based UI in PHP, but scanners written in C / C++
- Besides Web UI:
  - Command line utilities

Why SPDX in HD?

SPDX is for sharing metadata

What is SPDX?

- SPDX= Software Package Data Exchange
- Defines a format for exchanging metadata about software packages
  - Licensing
  - Copyrights
  - Extension for much more
  - Formats: RDF/XML, Tag-Value and Excel
  - Standardizing license identification
    - SPDX License List

More Information

- SPDX Homepage
  https://spdx.org/
- SPDX Standard description
  https://spdx.org/specifications
- SPDX License list
  https://spdx.org/license-list
- SPDX Example?
  - You can use FOSSology

What is this about? (Again)

You know these examples

- It is about licensing of software packages
- Licensing is expressed in files
- You would like to find licensing statements …
- …and summarize this in SPDX

ZIP

jakarta-oro-2.0.8.zip

```
<spdx:hasExtractedLicensingInfo>
<spdx:ExtractedLicensingInfo
rdf:about="http://spdx.org/licenses/Apache-1.1">
<spdx:licenseId>Apache-1.1</spdx:licenseId>
<spdx:extractedText>
<![CDATA[
The Apache Software License, Version 1.1
*
* Copyright (c) 2000 The Apache Software Foundation.  All rights
* reserved.
*
```

Why do we call it HD - High Definition?

SIEMENS

- **How do you find actually licenses?**
  … it depends on scanners, text finding implementations.

- **How do you deal with thousands of files and dozens of licenses?**
  … you will need overview.

- **What if automated scanners cannot determine the licenses?**
  … not only scanning, but also reviewing.

- **Found a new license text?**
  … you will need to extract the text.

- **Copyrights with review?**
  … fishing them out too and eliminate false positives.

- **Dual licensing**

  … does the scanner find it?

Multiple Scanners for License Relevant Texts

Maturity, Precision, Speed

- FOSSology is a framework for agents
- Three scanners as agents
  - **Nomos:** Keyword and regular exp
  - **Ninka:** An sentence-based approach
  - **Monk:** Text similarity comparison
- Deferent scanners have their strength
- Perform differently for different contexts
  - For standard licenses
  - For finding license relevant statements

Overview gives you control

Handling Large Packages

- Smaller packages (<200 Files) and some licenses are easy to handle
  - Maybe tool support is even not required
- However, thousands of files, many folders and several different licenses are not ideal candidates for manual
  - How to keep overview and focus?
- FOSSology aggregates by two directions
  - Histogram of **Licenses** to focus on particular licenses
  - **Folder** hierarchy aggregation allowing

Review Enhances Precision

Scanner cannot determine all

- Some license relevant statements require review
  - References to files
  - New (or unknown) licenses
  - Uncommon statements about licensing
  - Different GPL versions
  - Similar licenses with a twist
- Example:
  - *"This software is licensed as described in the file COPYING, which*
  - *you should have received as part of this distribution. The terms*
  - *are also available at http://subversion.tigris.org/license-1.html.*
  - *If newer versions of this license are posted there, you may use a*
  - *newer version instead, at your option.*

Custom License Texts

Licenses not "databased"

- License relevant statements can be
  - References to files
  - New licenses / unknown licenses
  - Uncommon statements about licensing
- Moreover, scanners can be imprecise
  - Different GPL versions
  - Different GPL exceptions
  - Undefined references

Copyright (c) 1996 L. Peter Deutsch: Permission is granted to copy and distribute this document for any purpose and without charge, including translations into other languages and incorporation into compilations, provided that the copyright notice and this notice are preserved, and that any substantive changes or deletions from the original are clearly marked.

Copyright Statements

Handling Copyrights

- An agent finds copyright relevant statements
- Copyright statement listing in a basic obligation in licenses
  - legal obligation
  - but also fair use
- Bulk edit of copyright statements
- Efficient sorting out of false positives
  - Corrections required with, for example:
    Gnomovision version 69,
    Copyright (C) year name of author

Home Search Browse Upload Jobs Organize Admin Help

fossology

Copyright/Email/URL/Author Browser

logout
User: fossy
Group: fossy

Version: [unknown], Branch: [unknown], Commit: [#unknown] unknown built @ 2016/08/24 18:49 UTC

Folder: Software Repository/
toybox-0.5.2.tar.gz

License Browser | File Browser | Copyright/Email/URL | ECC | Browse | License List | Search • View | Info • Refresh

Show all
Jump to: Emails | URLs | Authors or Maintainers
Show 10 ▾ entries   Search:

toybox-0.5.2.tar

Count ▾	Copyright	
2	Copyright 2012 Harvind Singh <harvindsingh1981@gmail.com>	✖
2	Copyright 2012 Andreas Heck <aheck@gmx.de>	✖
2	Copyright 2010 Rob Landley <rob@landley.net>	✖
2	Copyright 2006, 2007 Rob Landley <rob@landley.net>	✖
1	(c) yyunput(c, (yytext_ptr))	✖
1	(c) return c + 1; return name;	✖
1	copyrights to my work), then relaunched under a 2-clause BSD license in 2011. The following news entries predate that relaunch:</p>	✖
1	Copyright (C) 2008 David Anders <danders@amltd.com>	✖
1	Copyright (C) 2006 by Rob Landley <rob@landley.net>	✖
1	Copyright (C) 2006, 2013 by Rob Landley <rob@landley.net>	✖

Showing 41 to 50 of 109 entries   ◀ Previous Next ▶

Dual Licenses

SPDX OR Operator

- Multiple licenses for a file is common
  - One part is different than other part
  - AND-based listing of files
- But multi licensing is different
  - Copyright holder allows user for choosing under which license conditions the usage rights can be exercised
- Changes the situation:
  - GPL-2.0 AND BSD-3-Clause
  - GPL-2.0 OR BSD-3-Clause

FOSSology is Popular

List of Disclosed Users

- See the FOSSology.org website for the most recent list of disclosed users
- *Not on this list?* – Consider supporting the project with putting your logo there

FOSSology Conclusions

Most important

- **It is open source software**
  - No vendor lock-in, it can be shared among partners
- **It supports SPDX**
  - Supporting the leading standard for package description
- **It has features for detailed and advanced SPDX generation**
  - Several features resulting from evolution over years of license clearing

You like it and would like to contribute?

- Put your logo on the FOSSology homepage if you use it
- Report issues on Github: https://github.com/fossology/fossology
- Consider contributing to the code base with your extensions

Thank you for your attention … questions?

SIEMENS

Michael C. Jaeger
Siemens Corporate Technology

Otto-Hahn-Ring 6
81793 Munich, Germany

E-mail:
michael.c.jaeger@siemens.com

Internet
siemens.com/corporate-technology

More links
https://github.com/fossology/fossology
https://github.com/sw360/sw360portal