FOSSology: News and Advances from the Project

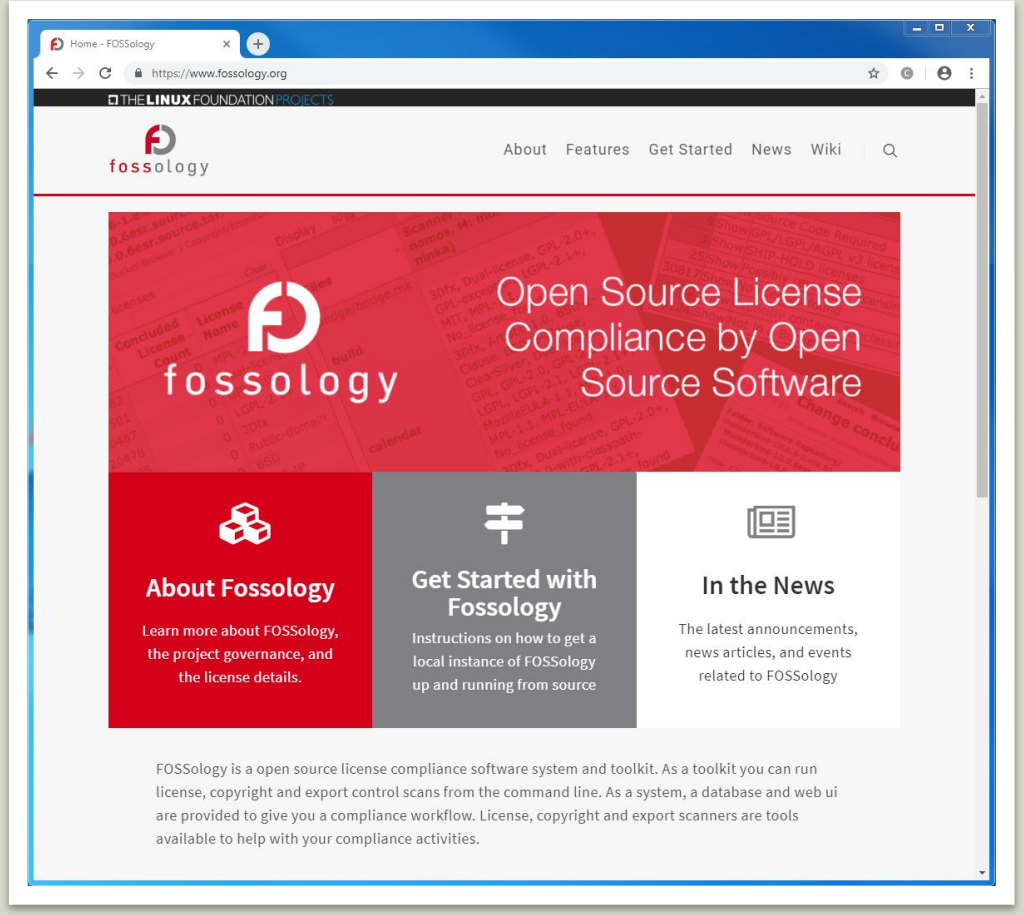
Presenters: Michael C. Jaeger, Siemens AG & Maximilian Huber, TNG Technology Consulting GmbH

CC-BY-SA 4.0

FOSSology – Linux Foundation Collaboration Project

www.fossology.org

- 2008 initial publication by HP
- 2015 Linux Foundation Collaboration Project
- It is a Linux Application
- Different tasks for OSS license compliance
  - Scanning for licenses
  - Copyright, authorship, e-mails
  - ECC statements
  - Generation of documentation
  - Export and import SPDX files



CC-BY-SA 4.0

FOSSology – It is about Overview

High Level and Drill Down

- Aggregation
  - Folder hierarchy of license findings
  - License-statement oriented view on files
  - Copyright aggregation
- Drill down
  - Navigate into folders
  - Filtering
  - Identify “the single” file

Display 25 licenses

Clear

Scanner Count	Concluded License Count	License Name
7702	8018	EPL-1.0
2339	52	Apache-2.0
275	0	MPL-2.0
112	0	MPL-1.1
110	0	LGPL-2.0+
110	0	Dual-license
64	0	Apache-possibility
57	23	W3C
51	50	MIT
49	0	GPL
34	0	W3C-IP
24	0	W3C-possibility
18	0	Public-domain
13	11	BSD-3-Clause
12	0	WebM
8	8	Apache-1.1
7	8	Apache-1.1-variant-jakarta-oro
6	0	CPL-1.0
5	0	W3C-style
4	0	UnclassifiedLicense
4	0	CPL-0.5
4	0	BSD-style
3	0	Microsoft-possibility
2	0	libtiff
2	0	Unicode

Showing 1 to 25 of 42 licenses Page 1 of 2

Hint: Click on the license name to search for where the license is found in the file listing.

Display 50 files (**tree view** or flat)

Files	-- filter for scan resul... Scanner Results (N: nomos, M: monk, Nk: ninka, I: reportImport)	-- filter for edited res... Edited Results
com.lowagie.text_2.1.7.v201004222200.jar	Adobe, Apache-2.0, APAFML, BSD-3-Clause, CUA-OPL-1.0, EPL-1.0, LGPL-2.0, libtiff, MIT-style, MPL-1.1, No_license_found, Permission Notice, Unicode	EPL-1.0, Apache-2.0
com.lowagie.text.source_2.1.7.v201004222200.jar	Apache-2.0, BSD-3-Clause, CUA-OPL-1.0, Dual-license, EPL-1.0, LGPL, LGPL-2.0+, libtiff, MIT, MIT-style, MPL, MPL-1.1, No_license_found, Permission Notice, Public-domain, Unicode, WebM	MIT, BSD-3-Clause, EPL-1.0
javax.wsdl_1.5.1.v201012040544.jar		
javax.xml.rpc_1.1.0.v201209140446.jar		Apache-2.0
javax.xml.soap_1.2.0.v201005080501.jar		Apache-2.0
javax.xml.stream_1.0.1.v201004272200.jar		Apache-2.0
org.apache.axis_1.4.0.v201411182030.jar	Apache-2.0, Apache-possibility, EPL-1.0, No_license_found, W3C-possibility	Apache-2.0
org.apache.batik.bridge_1.6.0.v201011041432.jar	Apache-2.0, Apache-possibility, EPL-1.0, No_license_found, Public-domain, W3C, W3C-IP	W3C, EPL-1.0, Apache-2.0
org.apache.batik.bridge.source_1.6.0.v201011041432.jar	Apache-2.0, Apache-possibility, EPL-1.0, No_license_found, Public-domain, W3C, W3C-IP	W3C, EPL-1.0, Apache-2.0

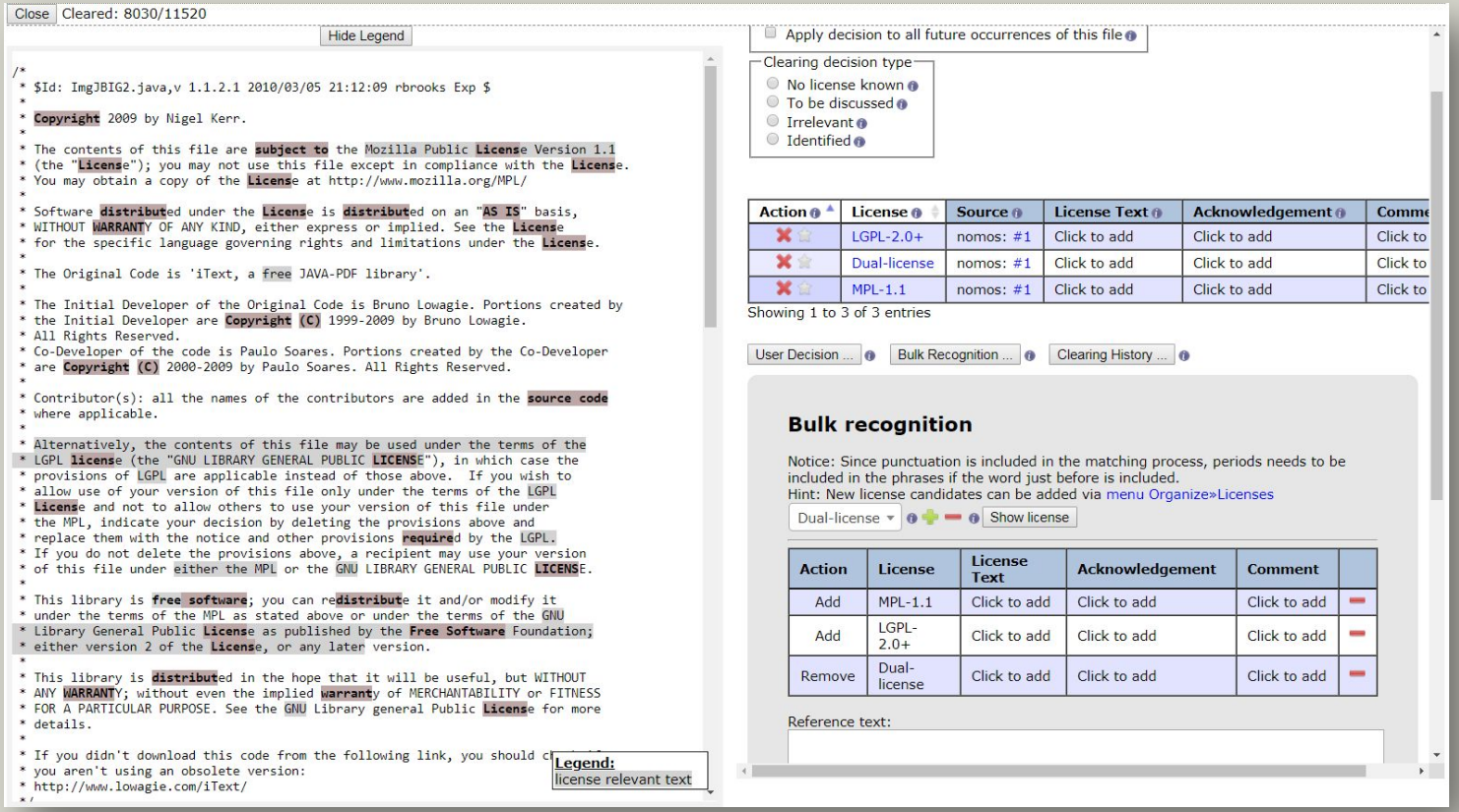
Recursive unpacking of files too!

CC-BY-SA 4.0

FOSSology – Review Findings

Specialized in Review

- Single file review
  - Highlighting of license relevant content
  - Reference text comparison
  - License statement decisions on statement level (“bulk scan”)



CC-BY-SA 4.0

FOSSology – It is about Conclusions

Licensing Challenges

- Licensing can be simple …
- … or challenging:
  - Unknown Licenses
  - Written statements
  - Unclear statements
  - Ambiguous statements
  - Incomplete statements
- Depends on domain
- Can be 30% hard to decide

```
SPDXVersion: SPDX-2.0
DataLicense: CC0-1.0
##-------------------------
## Document Information
##-------------------------
DocumentNamespace:
http://debian/repo/SPDX2TV_fossology-master-3.zip_1490661487.spdx
...
##File
FileName: fossology-master/utils/fo-installdeps
SPDXID: SPDXRef-item361
FileChecksum: SHA1: 3fc0aa4a4face8a0d317e0272c5e28e43f44c45a
FileChecksum: MD5: 1576b827a8b28ce1513a490fe2fecdcd
LicenseConcluded: GPL-2.0
LicenseInfoInFile: GPL-2.0
FileCopyrightText: <text> Copyright (C) 2008-2014 Hewlett-Packard
Development Company, L.P. </text>
…
```


CC-BY-SA 4.0

FOSSology SPDX Import and Export

Import = Consuming SPDX

- Consistency!
  - Handling SPDX conclusions
  - Handling copyright statements
  - Handling new licenses
- Goal was to consistently import the data given existing records

SPDX import is the real exchange

CC-BY-SA 4.0

FOSSology – SPDX and it‘s Import Use Cases

Review 3rd parties

- If you receive SPDX … how to check?
- Similar problem to reviewing scanner findings
- Importing SPDX description on an uploaded package shows the SPDX conclusions
- Can chow even along own scanner finding

Share analysis work

- Creation of compliance documentation is work intensive
- Organisations use different tools
- SPDX info can be shared between different tools

Reuse analysis work

- SPDX info of a component is available
- You need to generate SPDX info of a newer version of a component
- SPDX info can be imported and reused for files (based on hash) which did not change

CC-BY-SA 4.0

New Versions mean new Features

...

REST API
SPDX LL Updates
Obligations

3.5

2019-04-12

JSON and Lics
Ojo - SPDX Tags
REST API++

3.6

2019-09-11

Decider++
GCC8 Compliance
REST API++

3.7

RC 1 @
2019-10-24

...

CC-BY-SA 4.0

Ojo, a scanner to detect SPDX-License-IDs

```
#include<stdio.h>
/*
 * Written by John Doe
 * SPDX-License-Identifier: Apache-2.0
 */
int main() {
  printf("Hello World\n");
  return 0;
}
```

Gets detected as Apache-2.0

CC-BY-SA 4.0

Combine Ojo results with the other scanners

Automatize

- The Ojo information can be combined with the other findings
- If no other scanner found a contradicting statement, the result can be concluded

- [x] Monk License Analysis, scanning for licenses performing a text comparison
- [x] Nomos License Analysis, scanning for licenses using regular expressions
- [x] Ojo License Analysis, scanning for licenses using SPDX-License-Identifier
- [ ] Package Analysis (Parse package headers)

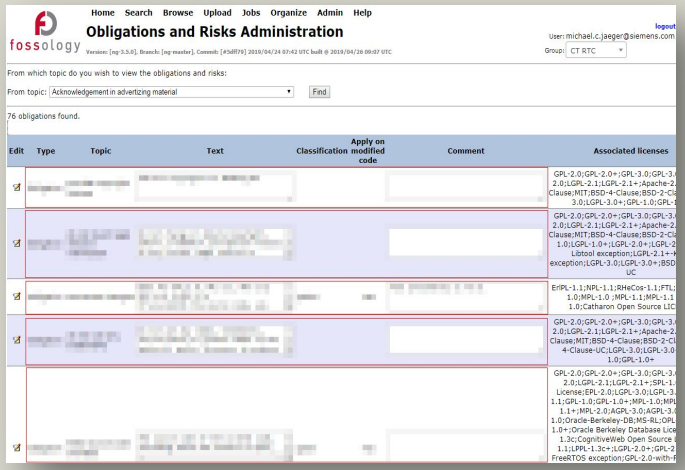
7. Automatic Concluded License Decider, based on
   - [ ] ... scanners matches if all Nomos findings are within the Monk findings
   - [x] ... scanners matches if Ojo findings are no contradiction with other findings
   - [ ] ... bulk phrases from reused packages

CC-BY-SA 4.0

FOSSology – License Obligations

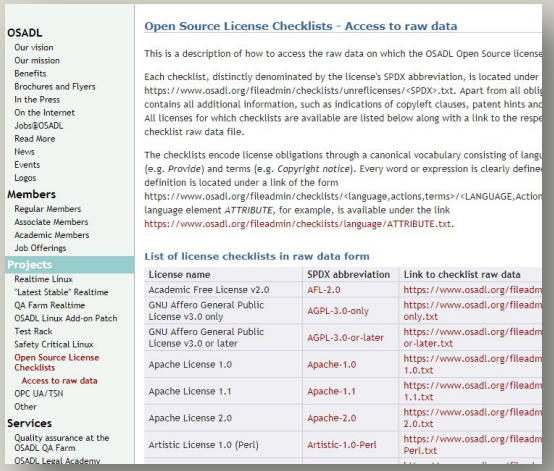
Obligation Mngmt

- Attach obligation entries to licenses
- Admin management UI
- Report documentation for components



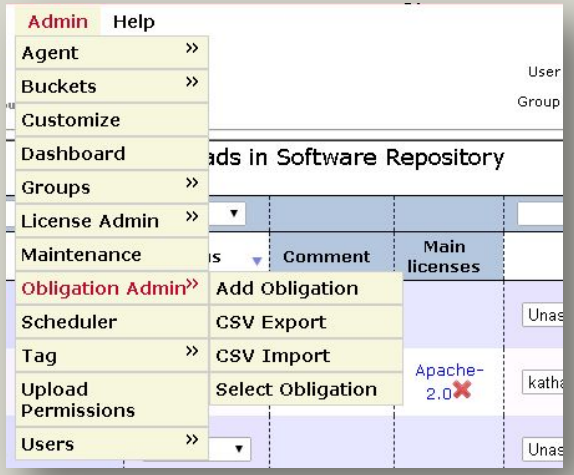
Obligation Source

- Different sources available
  - OSADL License Checklist
  - FINOS OSS Handbook
  - Github: Choose-a-license
- Machine readable formats



Obligation Import

- FOSSology can import records
- Currently: Convert your own data
- Potentially hosted conversion of obligations



CC-BY-SA 4.0

Integration

CC-BY-SA 4.0

FOSSology – Of course you can automate!

REST API

- Manage folders, uploads
- Trigger scans and options
- Download reports
- More info at: https://www.fossology.org/get-started/basic-rest-api-calls/
- (complete flow explained)

FOSSdriver

- Python based library
- Write your own Python workflow
- Not only what REST API can do
  - … but also manage bulk scans
- More info at: https://github.com/fossology/fossdriver

Command line tools

- Many functions and agents have command line interfaces
  - Nomos an Monk license scanners
  - Copyright scanner
  - License listings
  - …
- Upload and download tools

CC-BY-SA 4.0

FOSSology REST API – Very Straight Forward

Prepare

- List folders
- Create a folder if necessary
- Upload a package, OSS component to a folder

Scan

- Schedule scan jobs
- Set options for the jobs

Observe

- List running jobs
- Check their states

Download

- Download reports
  - SPDX
  - Word report
  - Readme, License Listing with all license texts and copyright statements

- More information: https://www.fossology.org/get-started/basic-rest-api-calls/
- Or use REST interface documentation: https://github.com/fossology/fossology/blob/master/src/www/ui/api/documentation/openapi.yaml

CC-BY-SA 4.0

Live Demo?

CC-BY-SA 4.0

How about machine learning?

CC-BY-SA 4.0

Machine Learning: From What Do You Learn?

Learning from Software vs. Learning from Humans

- Goal: Provide a License Classifier
- Sources for license classification data
  - Scanners in general for license statements
  - SPDX Identifiers
  - Particular implementations: nomos, monk, atarashi, scancode
- Programs will be only as good as programs
  - FOSSology is unique to this regards, because of its review interface

CC-BY-SA 4.0

FOSSologyML - Step 1 of 3

Create Model

- A database extractor accesses the FOSSology database
- Extracts license conclusions
- Along with the matched text
- Bulk matches become handy

- More info at: https://github.com/fossology/atarashi

CC-BY-SA 4.0

Text Processing: Cleanup and Lemmatization

```
#include<stdio.h>
/* Written by John Doe
 * This code is licensed under MIT
 */
int main() {
  // this is a simple hello world program
  printf("Hello World\n");
  return 0;
}
```

… will be transformed into:

```
Written by John Doe Thi code is licens under MIT
```


CC-BY-SA 4.0

FOSSologyML - Step 2 of 3

Create Model

- A database extractor accesses the FOSSology database
- Extracts license conclusions
- Along with the matched text
- Bulk matches become handy

- More info at: https://github.com/fossology/atarashi

Preprocess Files

- Preprocessing is necessary
- Otherwise too much clutter „confuses the learner“
- Strip out programming language parts (if recognized)
- Extract comments from files with programming language
- Lemmatization

CC-BY-SA 4.0

Machine Learning: Data You Get

Toxic for classifiers: many similar texts

- Goal: Provide a License Classifier
- Problem: Many very similar texts
- Regardless after lemmatization
- Example: one production server:
  - about 1000 license texts
  - about 4000 license statements
- Very similar:
  - BSD license variants: changes in wording w.r.t. author / copyright holders
  - MIT, NTP and similar simple permissive ones
  - FSF license variants: super simple one liners

CC-BY-SA 4.0

FOSSologyML - Step 3 of 3

Create Model

- A database extractor accesses the FOSSology database
- Extracts license conclusions
- Along with the matched text
- Bulk matches become handy

- More info at: https://github.com/fossology/atarashi

Preprocess Files

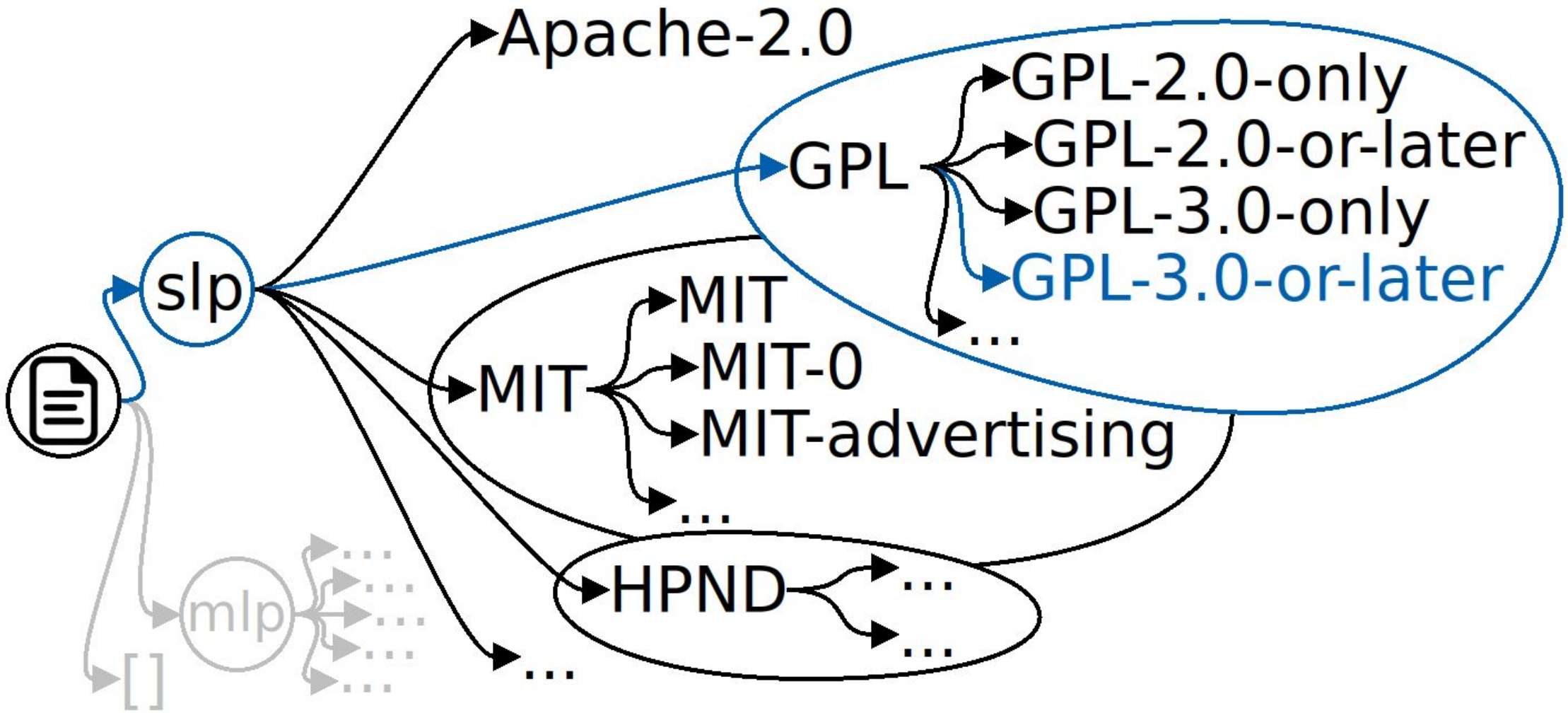
- Preprocessing is necessary
- Otherwise too much clutter „confuses the learner“
- Strip out programming language parts (if recognized)
- Extract comments from files with programming language
- Lemmatization

Staged Models

- A single model for all licenses did not show good performance
- Solution is to stage the recognition into sub problem
- By this approach, a single classifier determines only if a licensing statement is present or not.

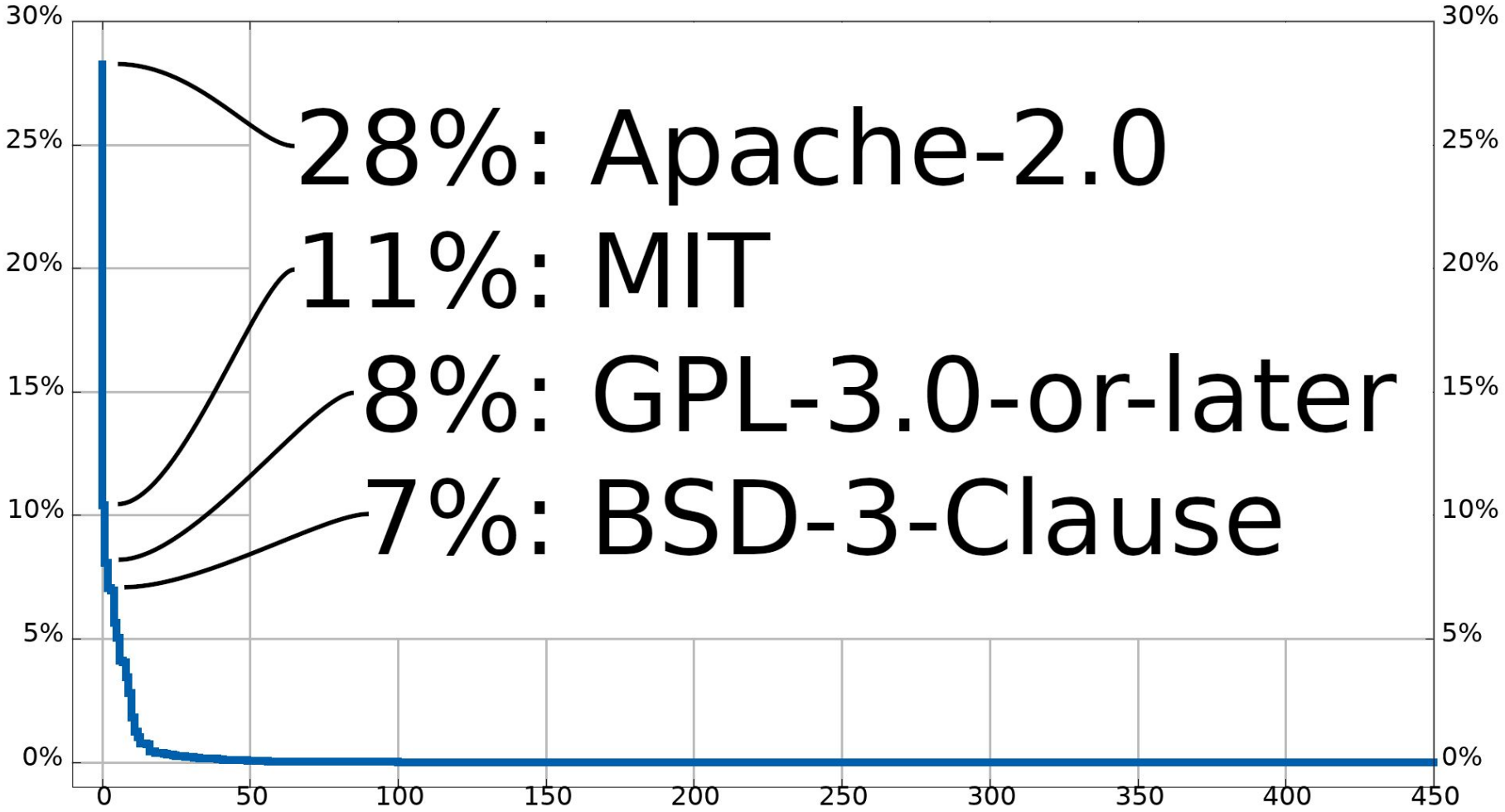
CC-BY-SA 4.0

Conquer and Divide



CC-BY-SA 4.0

Dilemma Biased Training Set



CC-BY-SA 4.0

FOSSologyML – And?

It is there: https://github.com/fossology/FOSSologyML

- *A FOSSology instance is required!*

- Python code to extract data from FOSSology server
- Python code to build a model
- Python code to use the model standalone
- Install FOSSologyML agent to an existing server

- *Experimental:* not part of the FOSSology main distribution

CC-BY-SA 4.0

OSS Community

CC-BY-SA 4.0

FOSSology – Did we tell you about GSOC 2019?

Software Heritage

- A world wide archive and public repository
- Covering all kinds of published software
- Goal is to establish interaction with SH REST API

Clearly Defined

- Licensing and other metadata about OSS components online
- Goal to establish first steps interaction with CD REST API

Atarashi Integration

- Atarashi should replace Ninka
- Both are license scanners
- Ninka was good, but is not maintained anymore

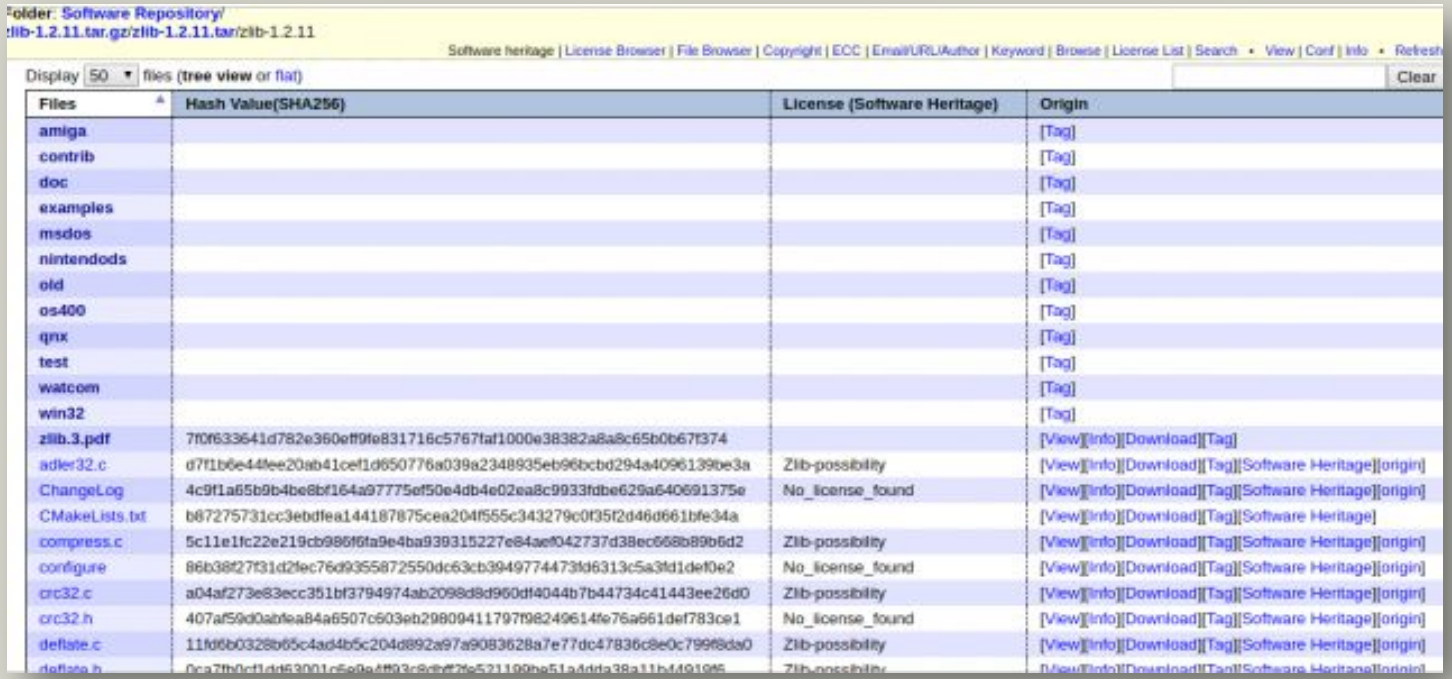
The FOSSology project was awarded three internship slots with the Google Summer of Code run in 2019! *(thank you so much Google, this is awesome!)*

CC-BY-SA 4.0

FOSSology and Software Heritage

A completely new use case

- Via REST API, software heritage can tell you if a file was published
- Boring for OSS components (Hopefully they have them in their archives)
- But good for
  - Detect changes in OSS
  - Non-OSS uploads



Files	Hash Value(SHA256)	License (Software Heritage)	Origin
amiga			[Tag]
contrib			[Tag]
doc			[Tag]
examples			[Tag]
msdos			[Tag]
nintendods			[Tag]
old			[Tag]
os400			[Tag]
qnx			[Tag]
test			[Tag]
watcom			[Tag]
win32			[Tag]
zlib.3.pdf	7f0f633641d782e360eff9fe831716c5767faf1000e38382a8a8c65b0b67f374		[View][Info][Download][Tag]
adler32.c	d7f1b6e44fee20ab41cef1d650776a039a2348935eb96bcbd294a4096139be3a	Zlib-possibility	[View][Info][Download][Tag][Software Heritage][origin]
ChangeLog	4c9f1a65b9b4be8bf164a97775ef50e4db4e02ea8c9933fdbe629a640691375e	No_license_found	[View][Info][Download][Tag][Software Heritage][origin]
CMakeLists.txt	b87275731cc3ebdfea144187875cea204f555c343279c0f35f2d46d661bfe34a		[View][Info][Download][Tag][Software Heritage]
compress.c	5c11e1fc22e219cb986f6fa9e4ba939315227e84aef042737d38ec668b89b6d2	Zlib-possibility	[View][Info][Download][Tag][Software Heritage][origin]
configure	86b38f27f31d2fec76d9355872550dc63cb39497744731d6313c5a3fd1def0e2	No_license_found	[View][Info][Download][Tag][Software Heritage][origin]
crc32.c	a04af273e83ecc351bf3794974ab2098d8d960df4044b7b44734c41443ee26d0	Zlib-possibility	[View][Info][Download][Tag][Software Heritage][origin]
crc32.h	407af59d0abfea84a6507c603eb29809411797f98249614fe76a661def783ce1	No_license_found	[View][Info][Download][Tag][Software Heritage][origin]
deflate.c	11fd6b0328b65c4ad4b5c204d892a97a9083628a7e77dc47836c8e0c799f8da0	Zlib-possibility	[View][Info][Download][Tag][Software Heritage][origin]
deflate.h	0ca7fb0cf1dd63001c5e9a4ff93c8dbff2fe521199ba51a4dda38a11b449195	Zlib-possibility	[View][Info][Download][Tag][Software Heritage][origin]

CC-BY-SA 4.0

There is more

Atarashi

- A novel license scanner using text statistics
- Goal is to drop in texts and let atarashi find it
- Can be run stand alone
- *Integration in FOSSology currently beta*

- More info at: https://github.com/fossology/atarashi

FOSSologySlides

- Slides for Presenting FOSSology

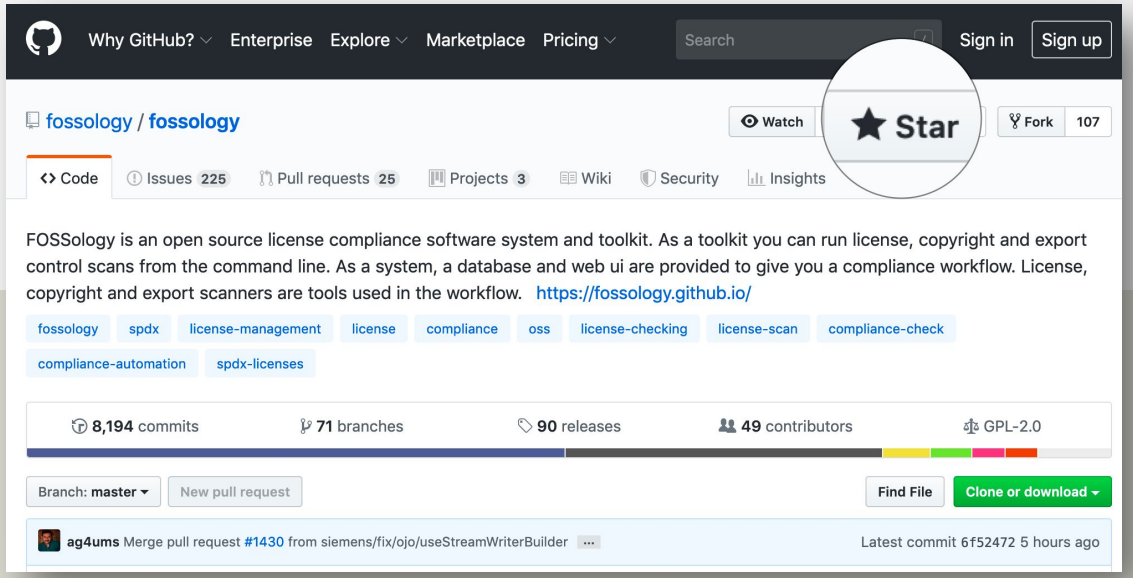
- More info at: https://github.com/fossology/FOSSologySlides

CC-BY-SA 4.0

Questions? – Consider to “Star Us”!

Michael C. Jaeger
Siemens AG
michael.c.jaeger@siemens.com

Maximilian Huber
TNG Technology Consulting GmbH
maximilian.huber@tngtech.com



FOSSology links
https://www.fossology.org/
https://github.com/fossology/fossology

SW360 links
https://sw360.github.io/
https://github.com/sw360/sw360portal

CC-BY-SA 4.0