

VOLKSWAGEN GROUP

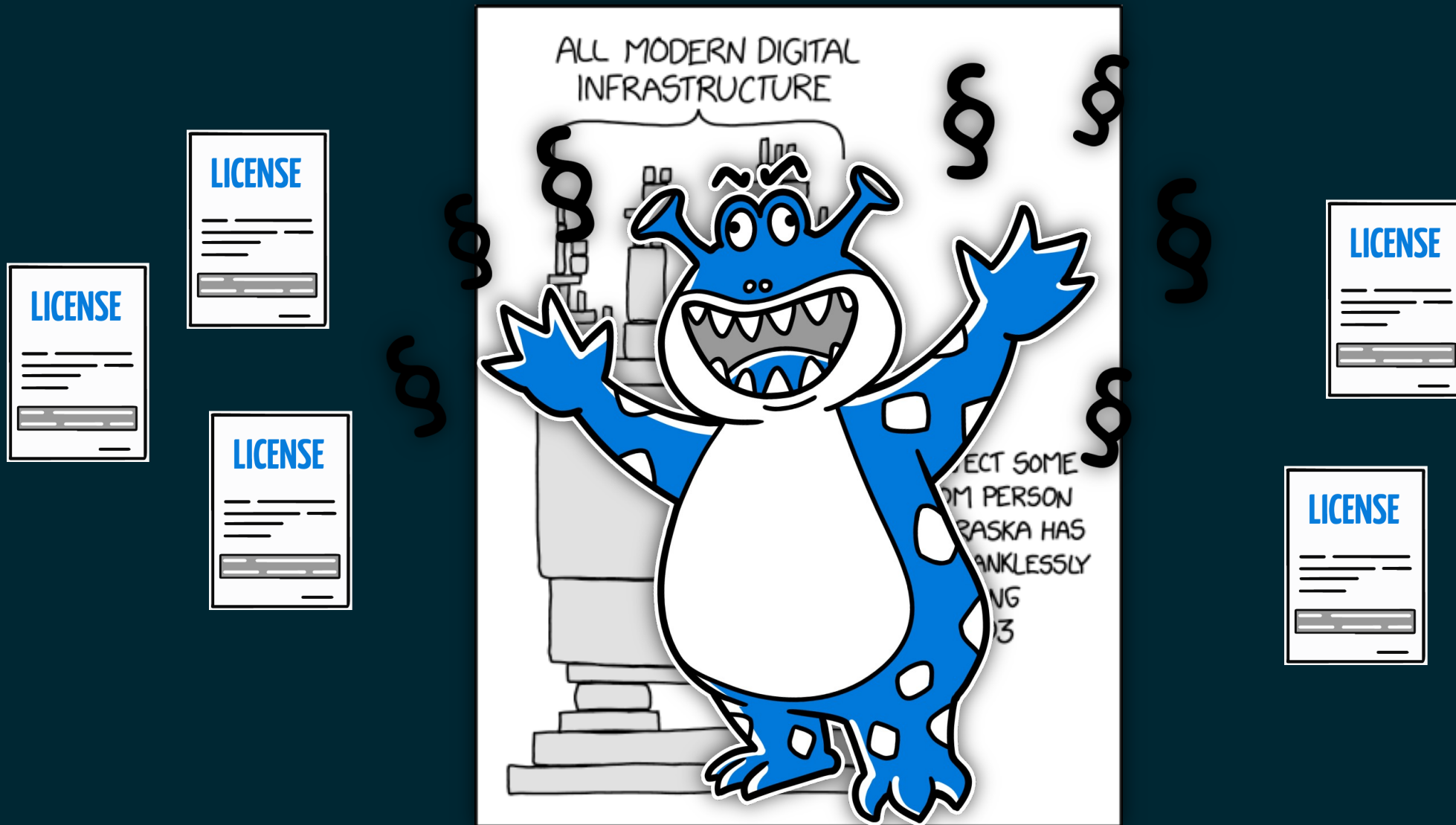


How Volkswagen uses ORT to build a curated database of software libraries

06.03.2024

PUBLIC





<https://xkcd.com/2347/>

Why build a curated database of software component licenses and copyrights?

OUR CHALLENGE

Strict requirements from our legal department:
complete + correct information

The old process..



Problems

- Lots of manual activities
- Media discontinuity
- Bad data in existing Database
- Teams are responsible to provide complete and correct data ⚡

The Product Vision

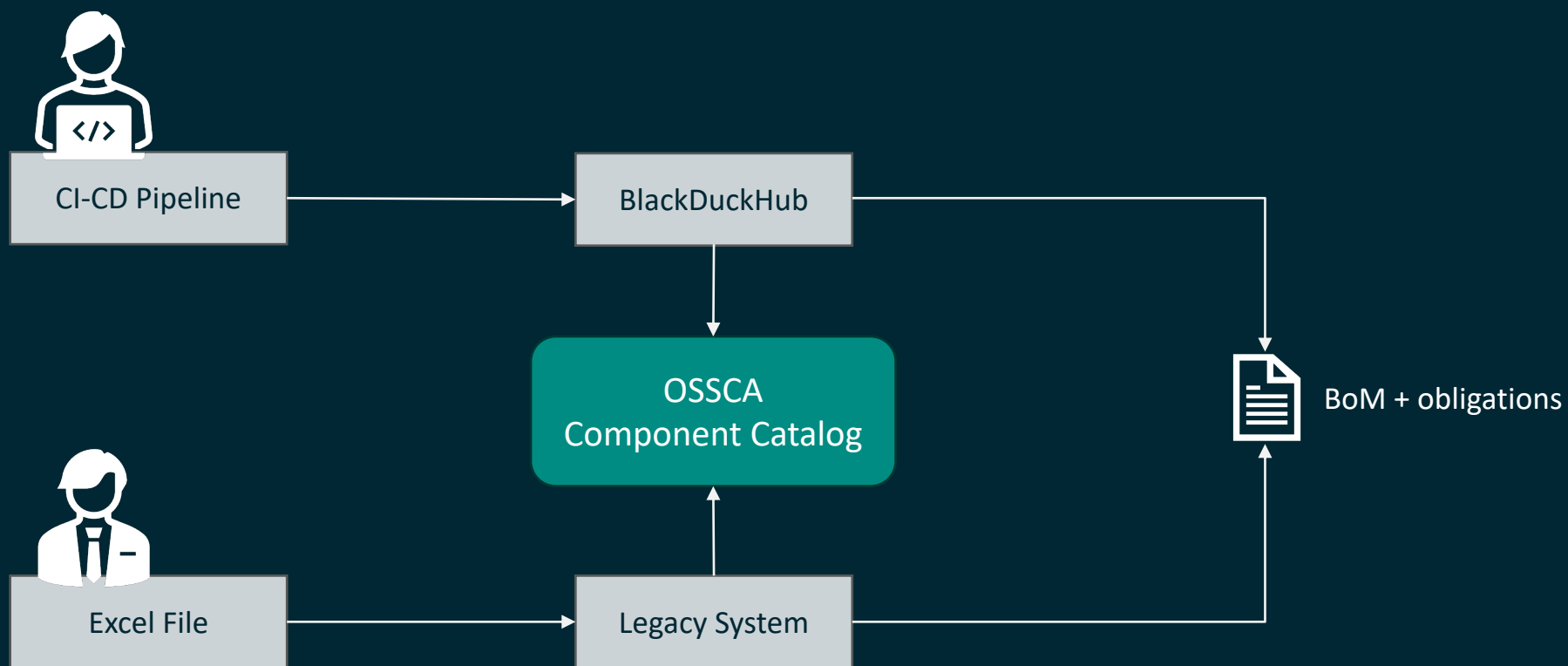
- Transfer data responsibility from developers to Open Source Office
- Open Source Office is service provider and ensures data quality in a curated database
- Curated data is used in an highly automated process

Why not use existing data?

- Incorrect / incomplete data
- Data ownership
- Time to fix incorrect entries

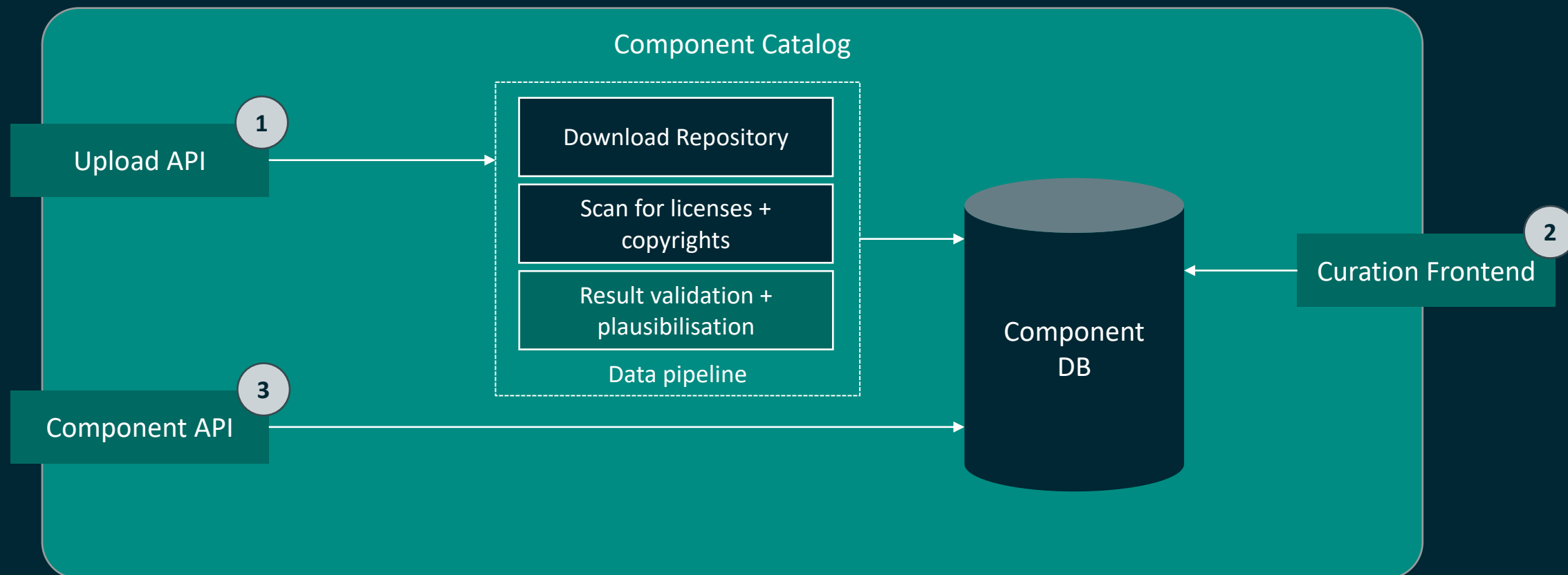
VW Open Source Software Clearance Assistant

A simple view



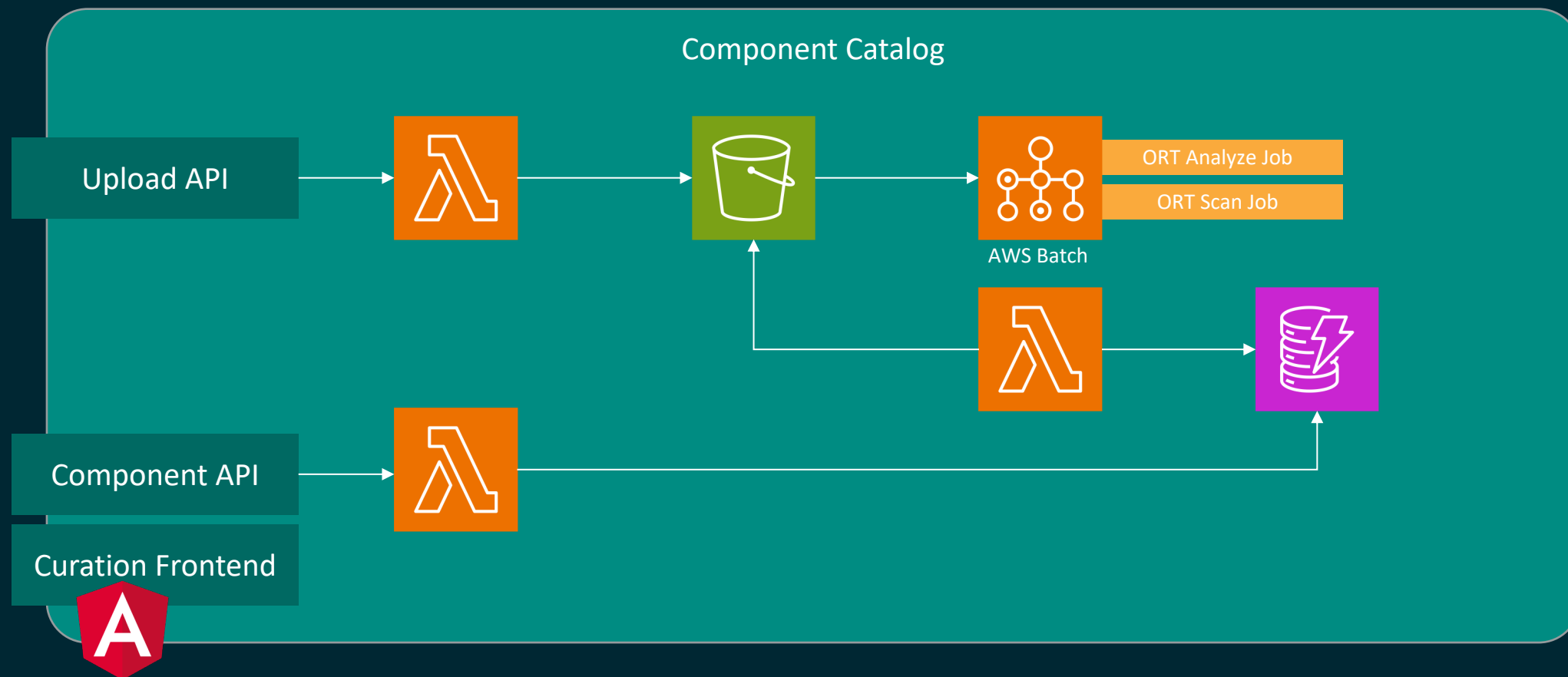
VW Open Source Software Clearance Assistant

A simple view



VW Open Source Software Clearance Assistant

Architecture in a nutshell



Validation + plausibilisation

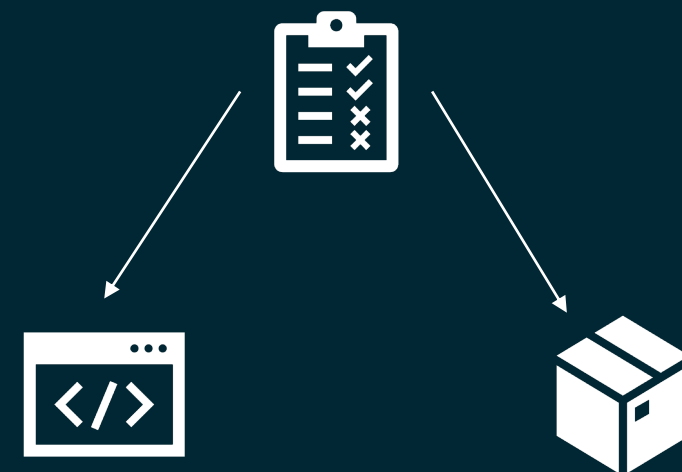
Why?

1. Ensure that declared source is decisive for associated release.
2. Ensure that selected repository revision is plausible for associated release.

Validation + plausibilisation

Example for NPM

- Compare published repository URL on NPM
 - Check if declared version is available on NPM
 - Evaluate content of publishing files
 - e.g. package.json, component.json, ...
 - Copyright checks (e.g. mark generic copyrights)
-
- Optimization: apply published repository directories in monorepos
 - Generic validations / errors are forewarded: like No revision candidate found
 - Assign curation priorities to files




Frontend Demo

Challenges

Data in the internet is... hard.

Challenges

Data in the internet is... hard.

- Copyright extraction from ScanCode is error-prone and does not hold the requirements of our legal department
 - Text is altered (e.g. copyright symbol)
 - Umlauts are dropped  → <https://github.com/nexB/scancode-toolkit/issues/1566>
 - Text junk

ScanCode result	Original line from file
Copyright (c) 1999-2004 Dmitriy Rogatkin	* Copyright (C) 1999-2004 Dmitriy Rogatkin. All rights reserved.
Copyright (c) 2015-2021 the original	# Copyright © 2015-2021 the original authors.
Copyright (c) 2003-2023 Sebastiano Vigna	* Copyright (C) 2003-2023 Sebastiano Vigna
Copyright (c) 2018 - 2023 a href https://github.com/facelessuser	copyright: Copyright © 2018 - 2023 Isaac Muse
Copyright 1998-2010 AOL Inc. - Apache	- PKCS#1 PEM encoded private key parsing and utility functions from oauth.googlecode.com - Copyright 1998-2010 AOL Inc. - Apache Commons Lang - https://github.com/apache/commons-lang
Copyright 2019 enforce-trailing-newline	"Copyright 2019", "enforce-trailing-newline"
Copyright The .NET project	Copyright ===== The .NET project copyright is held by ".NET Foundation and Contributors".

Challenges

Data in the internet is... hard.

- Extension to other fields besides the known development package managers is hard
 - Sources for linux package managers – Docker image clearance
 - Bad license / copyright situation in general
- Authors information as fallback for missing copyrights for correct attribution

Improve copyrights via ORT

Our problem statement and idea

Problem

ScanCode normalizes copyright information leading to information alteration or loss which leads to incorrect information which cannot be trusted in an automated way.

Solution(s)

1) Fix it in ScanCode

2) Add workaround to ORT

Compare identified copyrights against source information and add indicator of deviation with a low level of automated preprocessing.

Q&A