ClearlyDefined

Sharing is caring:
Open data for open source compliance

Philippe Ombredanne, AboutCode
Qing Tomlinson, SAP

Philippe Ombredanne

pombredanne@aboutcode.org
https://github.com/pombredanne
https://www.linkedin.com/in/philippeombredanne

- Lead maintainer of AboutCode
  - Open source code, data, and standards to automate and secure software supply chains
  - https://aboutcode.org
- Co-founder of ClearlyDefined
  - Member of technical steering committee
- Creator of PURL (Package-URL) and VERS, co-founder of SPDX, CycloneDX core contributor
- CTO and co-founder of nexB
  - Providing SCA services and AboutCode support since 2007
  - https://nexb.com

Qing Tomlinson

- SAP
  - Senior software developer
  - https://www.sap.com
- Maintainer of ClearlyDefined
  - Member of technical steering committee
  - https://clearlydefined.io/

qing.tomlinson@sap.com
https://github.com/qtomlinson

SBOMs Everywhere for Compliance and Security

The EU Cyber Resilience Act (CRA) and other regulations mandate comprehensive SBOMs, continuous vulnerability management, and supply chain transparency, creating significant operational and technical burdens.

Organizations face great challenges to generate SBOMs at scale for each stage on the supply chain, for every build or release, with accurate data.

Wasted Resources 🤦🏾‍♀️🤦‍♀️🤦🏼‍♂️

Organizations fix the same missing or wrongly identified FOSS package metadata and license over and over again. It is a waste of compute and human resources to rescan and reanalyze the same packages.

Duplicated compliance efforts should instead be coordinated, shared, and reused to improve efficiency for everyone.

FOSS for FOSS

Many proprietary or commercial databases (claim to) offer accurate metadata for software packages.

Data about open source packages must also be open. Anything else would be plain crazy dumb.

Community Solutions for Software Metadata

1. ClearlyDefined
   - Keyed by Coordinates (predates PURL)
   - Working together with AboutCode 🤝
   - Data: CC0 license
   - Code: MIT license
   - Centralized API
2. PurlDB (AboutCode)
   - Keyed by PURL
   - Joining forces with ClearlyDefined 🤝
   - Data: CC-BY or CC-BY-SA license
   - Code: Apache-2.0 license
   - Federated, eventually decentralized
3. deps.dev (Google)
   - Keyed by PURL
   - Data: Google ToS proprietary license
   - Code: some open source, some proprietary license
   - Centralized API
4. ecosyste.ms
   - Keyed by PURL
   - Data: CC-BY-SA license
   - Code: AGPL-3.0 license (throttled API) or commercial license
   - Centralized API

Other approaches: OSADL OSSelot, ORT Corrections and Curations, Libraries.io

ClearlyDefined

Bringing clarity to Open Source Software metadata and beyond.


Use the Data


Curate Data


Contribute Data


Contribute Code


Add a Harvest


Adopt Practices

ClearlyDefined Clearly Explained 🤔

1) Crowdsourced licensing metadata for every software component ever in a global database published for all to use

2) Cached copy of licensing metadata for each component with a simple API for integration and automation

3) Organizations contribute back with any missing or wrongly identified licensing metadata to improve accuracy

4) In a trusted non-profit: Open Source Initiative

Overview of ClearlyDefined Curation Process

website for ClearlyDefined
Purpose: Simplified view of definitions

human created PR

create PR

API PATCH /curations

service for ClearlyDefined
Purpose: Facilitate curating licenses
Purpose: API for accessing curations

clearlydefined/curated-data
Purpose: Repository of curations

Notify
PR merged

Write curation

curated-data
store

Merge PR

Review PR

human review

slide courtesy of E. Lynette Rayle (GitHub)

Vidal, N., Rayle, E. L., & Tomlinson, Q. (2024). *ClearlyDefined: A Crowdsourced Database of Licensing Metadata* [PowerPoint slides]. SOSS Fusion 2024, Atlanta, Georgia, United States. https://opensource.org/wp-content/uploads/2024/11/ClearlyDefined-SOSS-Fusion-2024.pdf

ClearlyDefined Benefits

- More accurate metadata
  - Multiple reviewers improve data quality
  - Corrections pushed upstream
- Sharing = Resource savings
  - Shared compliance efforts eliminates the need to rescan and reanalyze the same packages
  - SAP reported 30-50% reduction in review turnaround time after adoption
    - https://opensource.net/clearlydefined-at-sap/
- Forever open at open source initiative®

Organizations

Microsoft, GitHub, Cisco, SAP, Bloomberg Engineering, kusari, Eclipse Foundation, HPE, The Linux Foundation, OpenRail Association

Ecosystem

OSS Review Toolkit, OpenChain, AboutCode, SBOM, SPDX, CycloneDX, TODO, fossology, SW360, GUAC, FOSSLight, OWASP

DEMO

https://clearlydefined.io

Problems: Scale ⛰️🌎🌌

1) Growing volume of data (70TB) can make it difficult to use ClearlyDefined on-premises

2) Also challenges with scanning at scale, because ScanCode is super fast! 🤣

3) Diagnostics and debugging on a distributed application is hard 😅

Problems: Size...
55M+ packages

🪨🏋️‍♂️🪨🏋️‍♀️🪨🏋️

10,000+ curations!

https://clearlydefined.io/stats

55,196,539
Number of total definitions

60 Median licensed score
100 Median described score

npm	22,278,880 Total	60 Licensed	30 Described
gem	1,048,935 Total	62 Licensed	30 Described
pypi	4,631,550 Total	60 Licensed	100 Described
maven	6,371,686 Total	60 Licensed	100 Described
nuget	5,840,912 Total	45 Licensed	30 Described
git	4,343,197 Total	62 Licensed	100 Described
crate	87,500 Total	64 Licensed	100 Described
deb	394,262 Total	4 Licensed	100 Described

Problems: Infrastructure

1) Database license switcheroo necessitates migration from MongoDB to true FOSS alternative like DocumentDB

2) Migrate to federated data for performance and digital independence: unlock the data! 🔓

Problems: Data Contributions

1) Need to fix curation UI 🤢 to facilitate more contributions

2) Need more community contributors 👫🏋️‍♀️👫🏋️‍♂️🧑‍🤝‍🧑🏋️‍♀️👬

Future (2026) Plans

1) PURL support - New PURL API to query

2) Unlock the data - Distributed and federated reuse

3) Share all the scans, never scan twice the same package

4) New curation UI

5) All the SBOMs for all the packages

6) Establish close collaboration with AboutCode

Future (2027 and Beyond) Plans: Data Domination

All the data are belong to you. 😜

1) Expand ClearlyDefined to include all the pillars of data for compliance. We already have provenance and license, but add security and project health/lifecycle.

2) ClearlySecured (cybersecurity curations) for the vulnerabilities use case.

Join the Community!

1) Weekly developer meetings:

https://docs.clearlydefined.io/docs/community/meetings

2) Hang out in Discord:

https://discord.gg/wEzHJku

3) Contribute code and data:

https://docs.clearlydefined.io/docs/get-involved

4) Help sustain and grow ClearlyDefined!

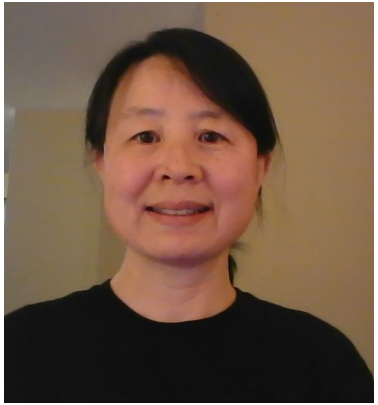
Contact pombredanne@aboutcode.org and qing.tomlinson@sap.com for more information.

Any Questions?

pombredanne@aboutcode.org
https://github.com/pombredanne
https://www.linkedin.com/in/philippeombredanne

https://clearlydefined.io

ClearlyDefined



qing.tomlinson@sap.com
https://github.com/qtomlinson

Get started with ClearlyDefined

https://docs.clearlydefined.io

Use the data

API: definitions, curations, harvest, attachments, notices

curl -X GET "https://api.clearlydefined.io/definitions/npm/npmjs/-/lodash/4.17.21" -H "accept: */*"

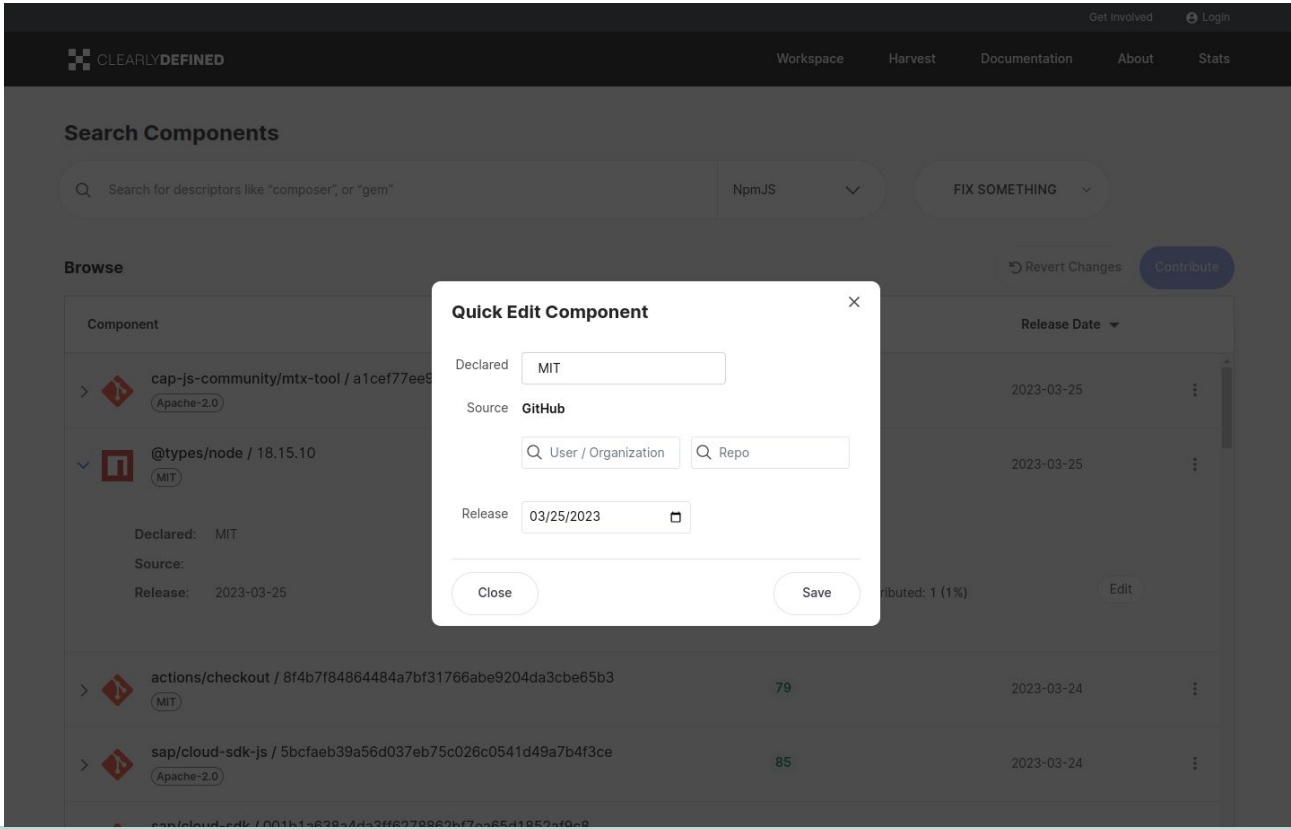
https://api.clearlydefined.io/api-docs/

Curate Data

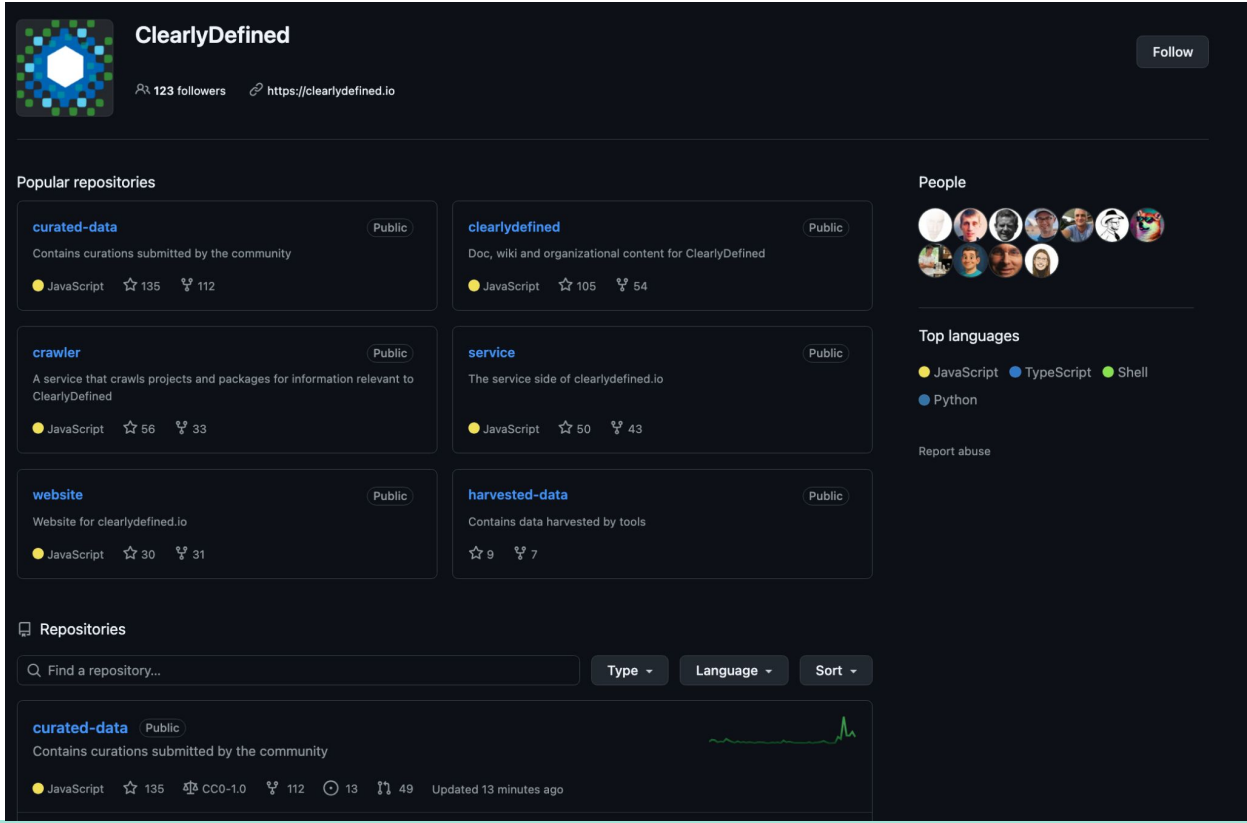
```
"contributionInfo": {
      "summary": "[Test] Update declared license",
      "details": "The declared license should be Apache as per the LICENSE file.",
      "resolution": "Updated declared license to Apache-2.0.",
      "type":"incorrect",
      "removeDefinitions":false
  },
```

https://docs.clearlydefined.io/docs/get-involved/data-curation

Contribute Data



Contribute Code



Add a Harvest

55,196,539
Number of total definitions

60 Median licensed score
100 Median described score

npm	22,278,880 Total	60 Licensed	30 Described
gem	1,048,935 Total	62 Licensed	30 Described
pypi	4,631,550 Total	60 Licensed	100 Described
maven	6,371,686 Total	60 Licensed	100 Described
nuget	5,840,912 Total	45 Licensed	30 Described
git	4,343,197 Total	62 Licensed	100 Described
crate	87,500 Total	64 Licensed	100 Described
deb	394,262 Total	4 Licensed	100 Described