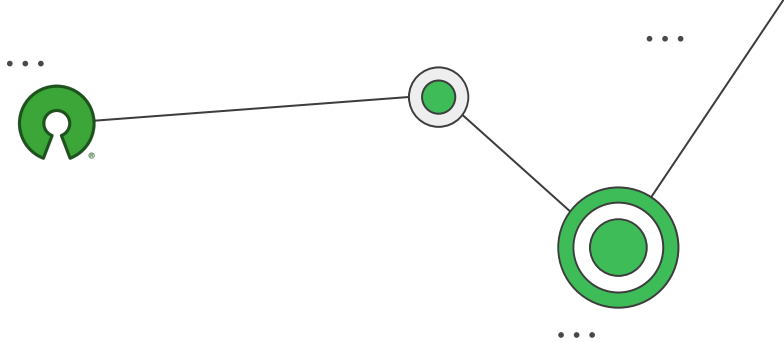


Sharing is caring: Open data for Open Source compliance

Nick Vidal, Alyssa Wright, Brian Duran

About us

ClearlyDefined Outreach Committee



Nick Vidal
OSI



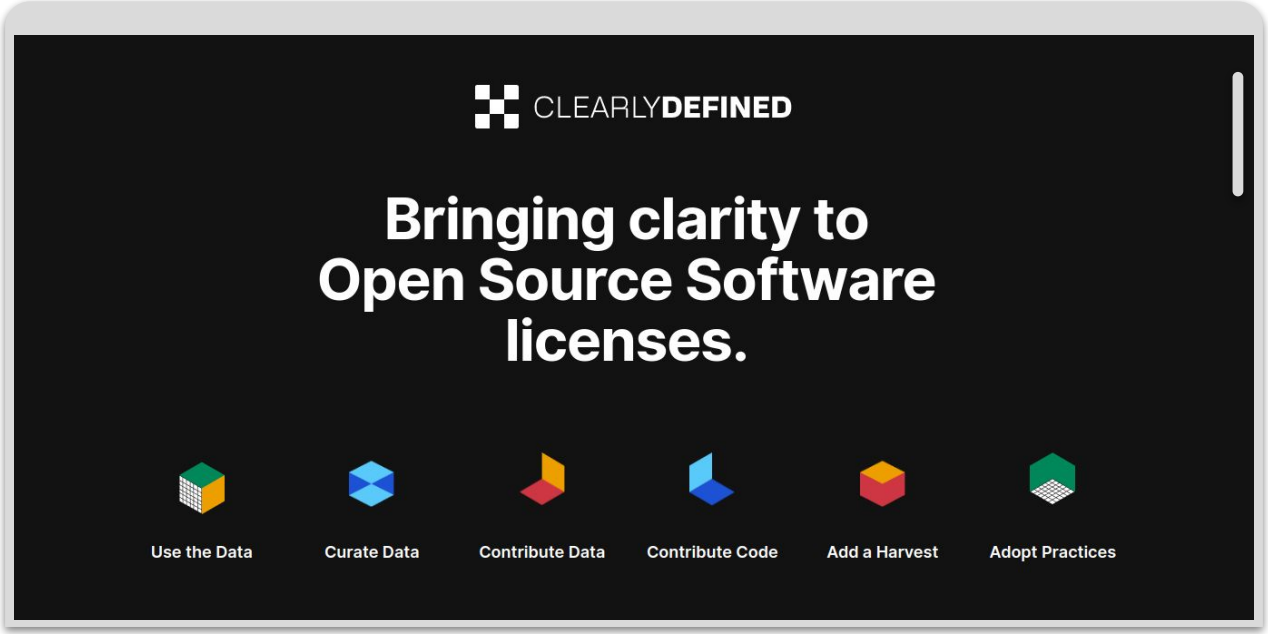
Alyssa Wright
Bloomberg



Brian Duran
SAP

ClearlyDefined

An open, crowdsourced database of license and attribution metadata for published open source.



ClearlyDefined

Bringing clarity to Open Source Software metadata and beyond.


Use the Data


Curate Data


Contribute Data


Contribute Code


Add a Harvest


Adopt Practices



Problem

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.



Problem

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.



Solution

This is where ClearlyDefined comes in, by serving a cached copy of licensing metadata for each component through a simple API for integration and automaton.

Organizations will also be able to contribute back with any missing or wrongly identified licensing metadata, helping to create a database that is accurate for the benefit of all.



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 the 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
 },
```

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 the OSI

open source initiative®

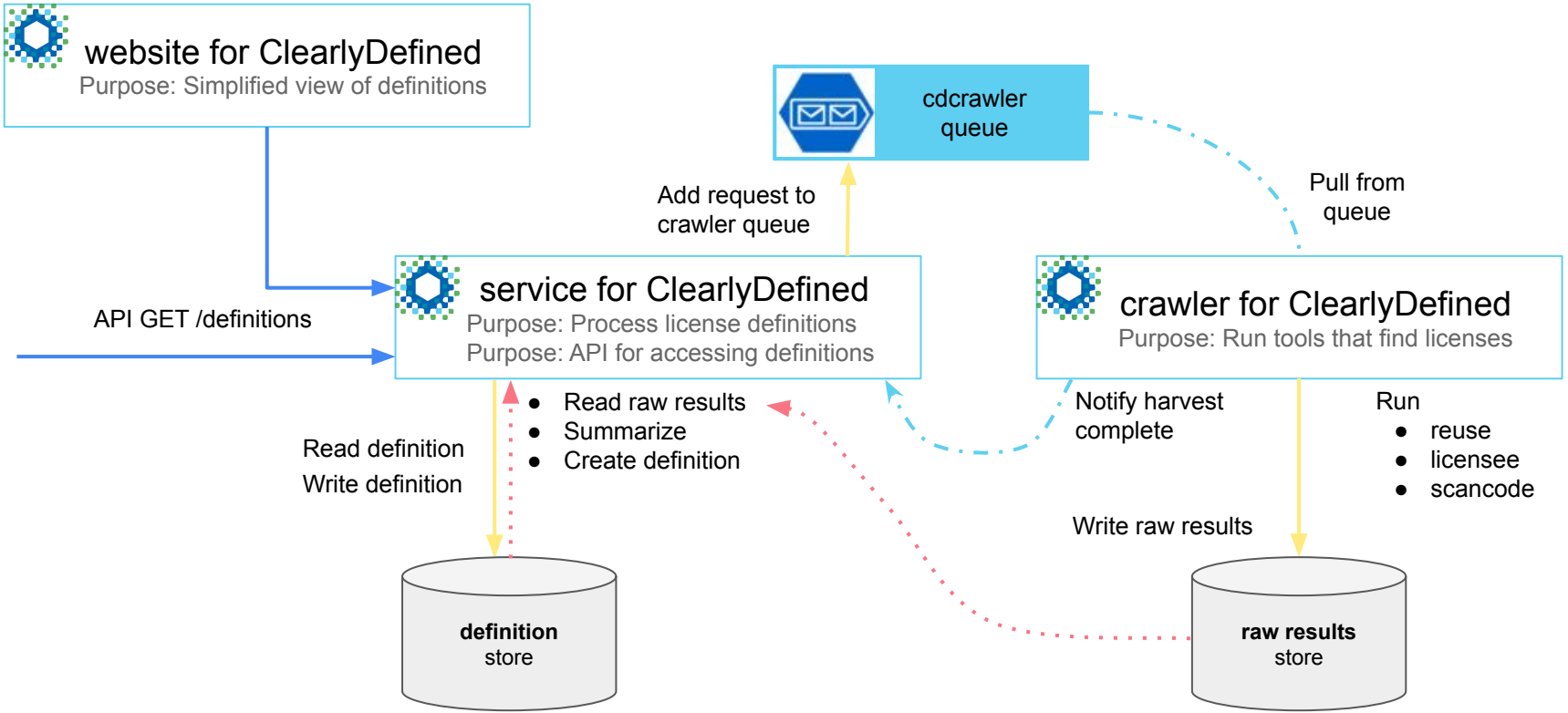
Organizations

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

Overview of ClearlyDefined Harvesting Process



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



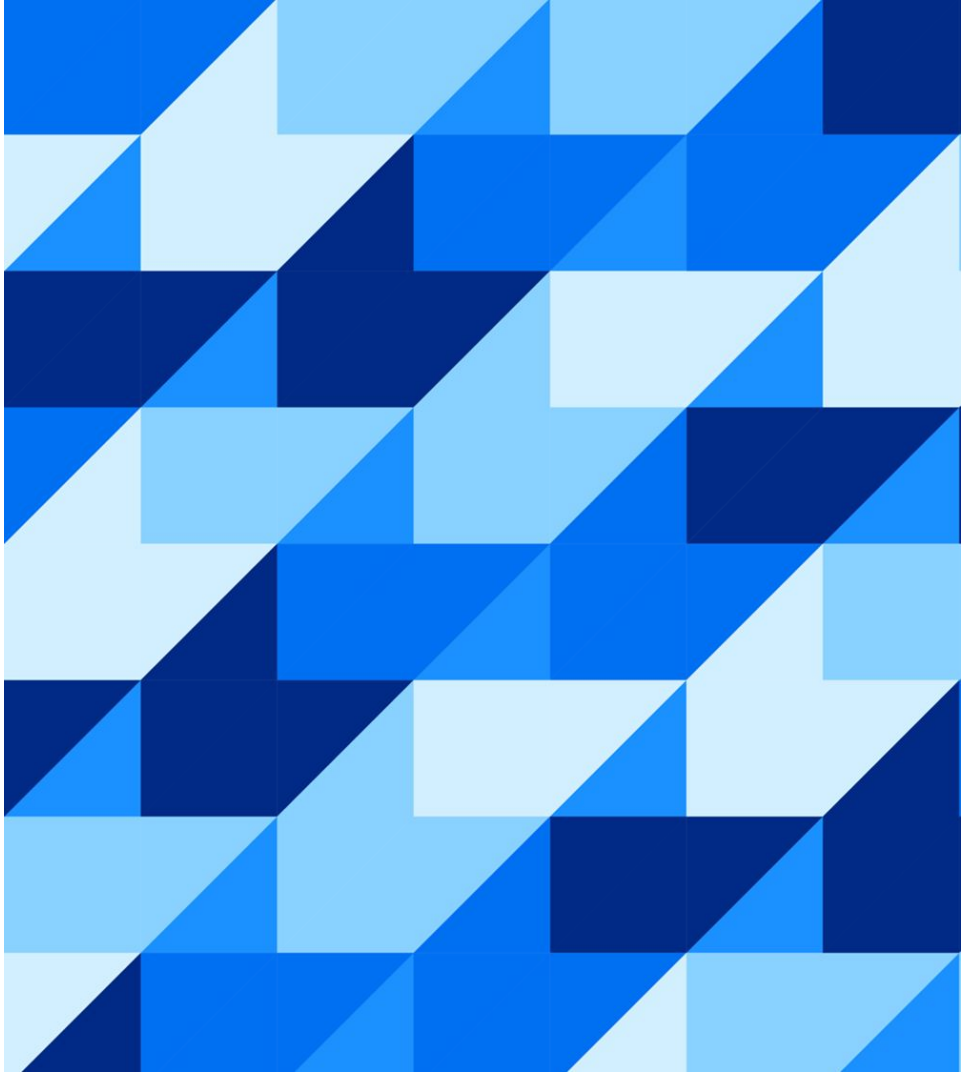
ClearlyDefined: LF case study

THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
NORTH AMERICA

Using SBOMs for Linux Foundation Projects

- LF adopting ClearlyDefined
  - Initially covering projects from CNCF, OpenSSF, and LF AI & Data, among others
  - Eventually will encompass all 1200+ LF projects
- Watch the recording: youtube.com/watch?v=u1_EJnoidYU





ClearlyDefined Adoption @SAP

SAP Open Source

We believe in co-innovation and collaboration with open source.



SAP's Open Source Vision

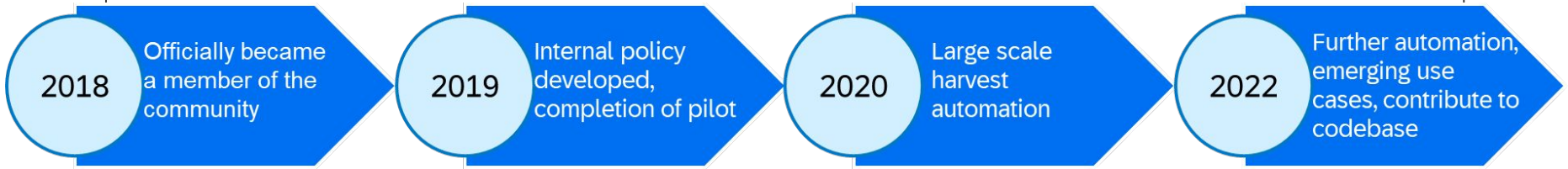
Driving innovation through open collaboration

Empowerment through openness

Trust and transparency

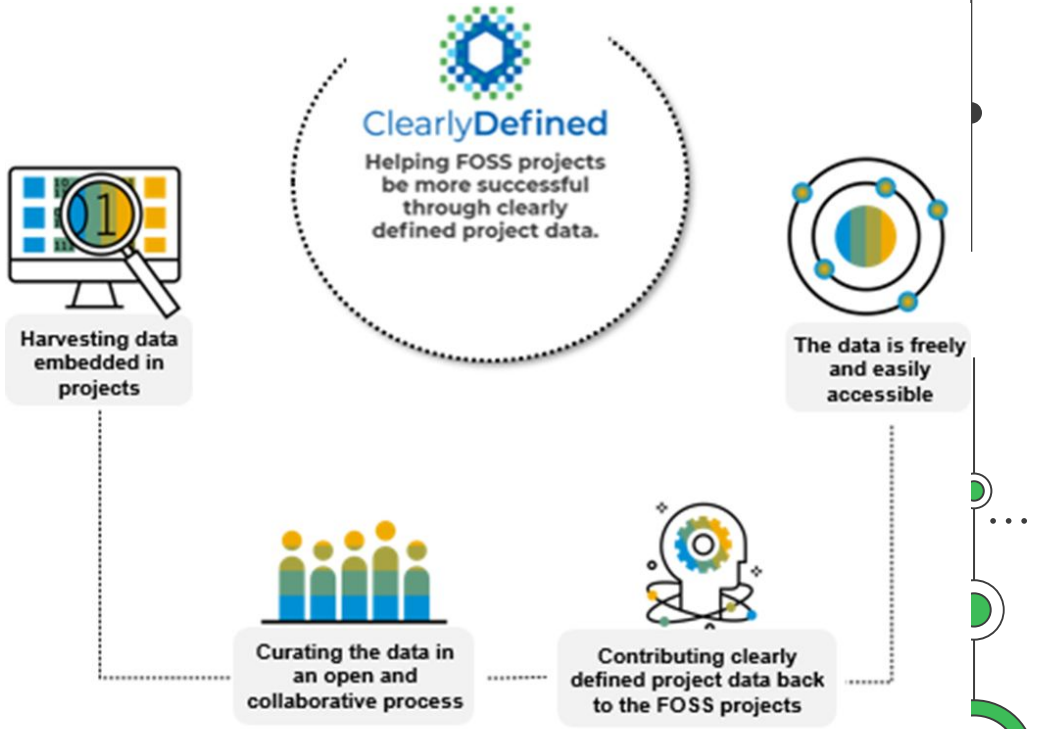
Community engagement

ClearlyDefined Adoption – Key Milestones

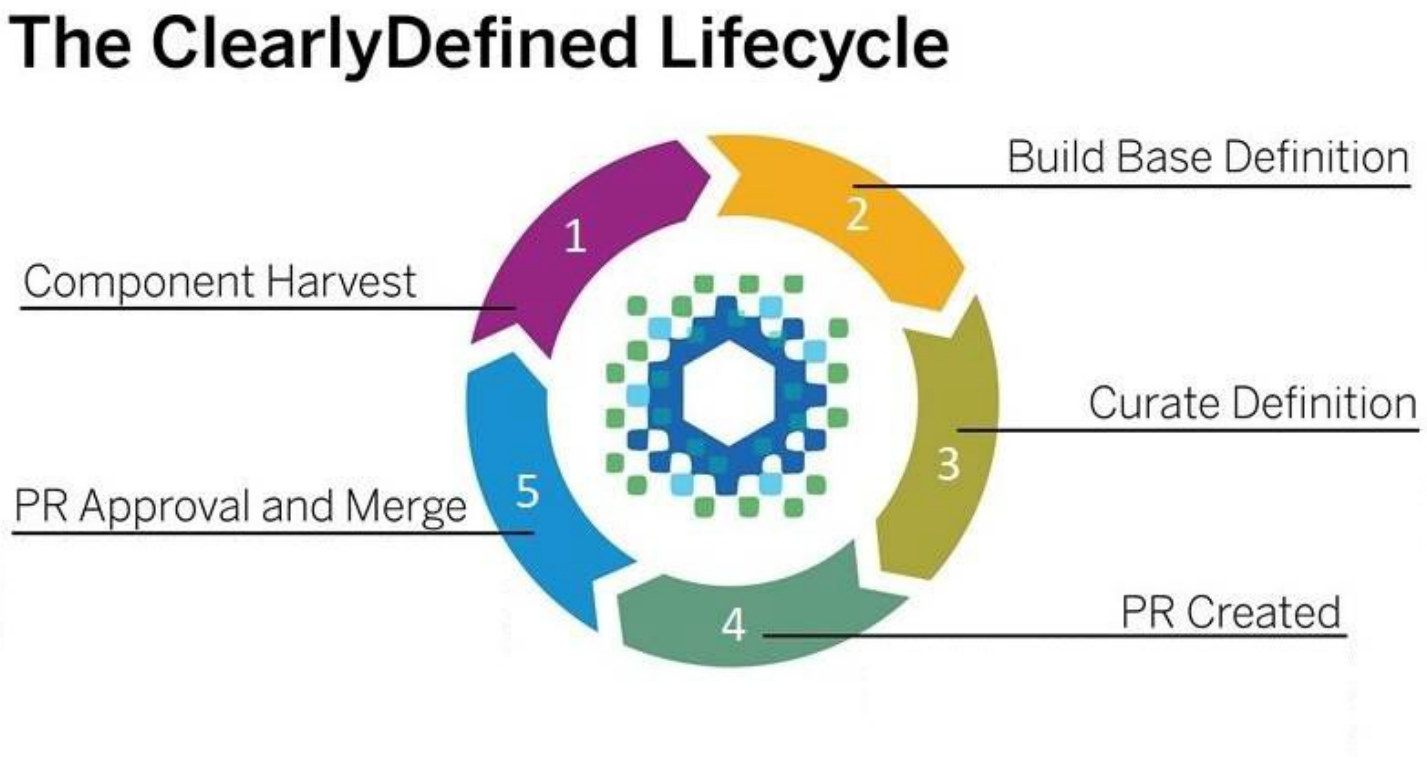


Why ClearlyDefined?

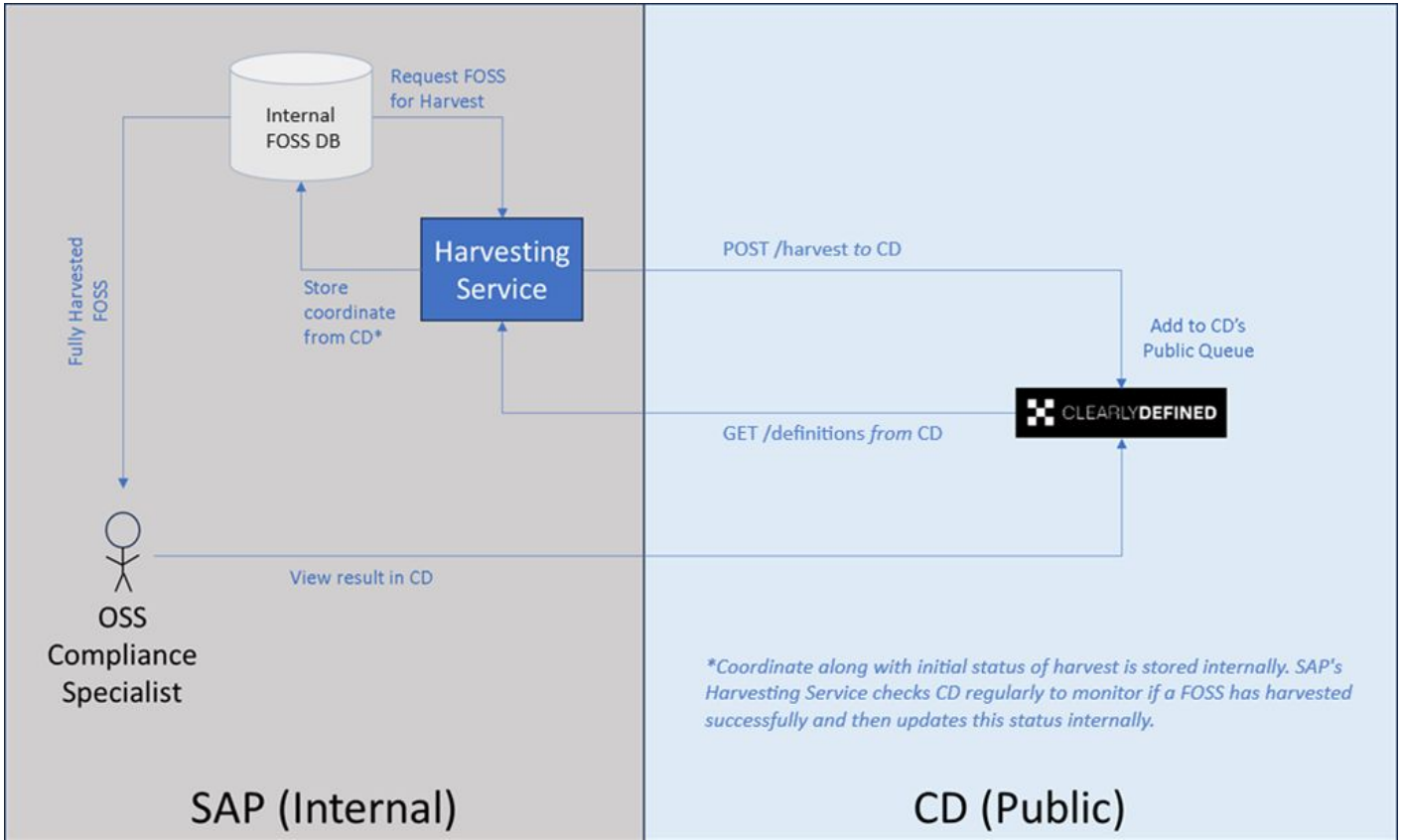
- ClearlyDefined supports an impressive and growing number of harvest sources
- Crowd-sourcing license review process saves time and effort



Giving Back – Contributions to Data



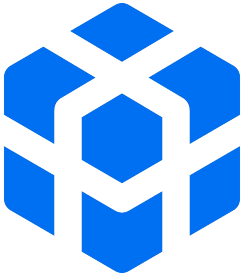
Overview of Automatic Harvesting



ClearlyDefined **Benefits**



Less time scanning, more time focusing on compliance



Better data quality



Community centered

Giving Back – Contributions to Code



Harvest Gradle Plugin



Optimizations



Analyses and Improvements



Collaborations

Key Takeaways

30-50% Improvement*



Community



*compared to the legacy (partly manual) process

ClearlyDefined: accomplishments

Version 2.0 released

We launched ClearlyDefined v2.0, a major milestone in improving license data quality

New documentation website

We unveiled a new website featuring comprehensive documentation and resources.

Open Governance

We elected leaders to the Steering and Outreach Committees.

Support for LicenseRef

By integrating support for LicenseRefs, we expanded beyond the SPDX License list.

Harvester for Conda

We introduced a new harvester for Conda with a focus on ML and data science packages.

Integration with GUAC

GUAC users now have access to enriched data for compliance and security.

ClearlyDefined: accomplishments

Conferences

We presented ClearlyDefined at conferences around the world (US, EU, APAC).

Ecosystem collaboration

We collaborated with adjacent communities: ScanCode, GUAC, and OpenChain.

SPDX and OWASP

We collaborated with SPDX and OWASP on licensing data specification.

New tools (cdsbom)

We developed new tools to enhance SBOMs with ClearlyDefined data.

New users

Organizations with wide reach like GitHub and Linux Foundation have adopted ClearlyDefined.

Secure Open Source

ScanCode first selected to GitHub's Secure Open Source Program.

ClearlyDefined: challenges

Infrastructure costs Costs to run infrastructure are high and funded by Microsoft.	**Technical debt** Implementation needs refactoring and is tightly integrated with Azure.	**Big investment** Investments to modernize the codebase are needed.
Scale Growing volume of data (70TB) can make it difficult to use on-premises.	**Access to data** Access currently is rate limited and there's a need for federated data.	**User Interface** The user interface needs an overhaul to make it easier for fixing licensing issues.

ClearlyDefined: MoU with AboutCode

New partnership

AboutCode, a public-benefit, non-profit organization, has become a key partner.

Leadership

Co-founders of ClearlyDefined and SPDX and contributors to other projects like CycloneDX.

Expertise

Creators of ScanCode and PURL, with recognized license compliance expertise.

Cost effective

They are in touch with stakeholders demonstrating how to save resources.

Sustainability

Organizations have expressed interest in providing technical and financial support.

Open governance

AboutCode supports an open governance model with community and the OSI.



Roadmap (2026-2028)

https://docs.clearlydefined.io/docs/resources/roadmap

1) Create better license and package scans

2) Reduce technical debt

3) Shared scans for shared costs

4) Integrate PURL and SBOM for easier interoperability

5) Improve curation processes and user experience

6) Support for more use cases

© Open Source Initiative / ClearlyDefined / CC-BY-4.0 / clearlydefined.io



Join the Community!



1) Weekly developer meetings:

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

2) Hang out in Slack:

https://aboutcode-org.slack.com/

3) Contribute code and data:

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

4) Help sustain and grow ClearlyDefined!

Contact nick@opensource.org for more information.

© Open Source Initiative / ClearlyDefined / CC-BY-4.0 / clearlydefined.io



THANKS!

CLEARLYDEFINED

Bringing clarity to Open Source Software licenses.



Use the Data



Curate Data



Contribute Data



Contribute Code



Add a Harvest



Adopt Practices

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: */*"

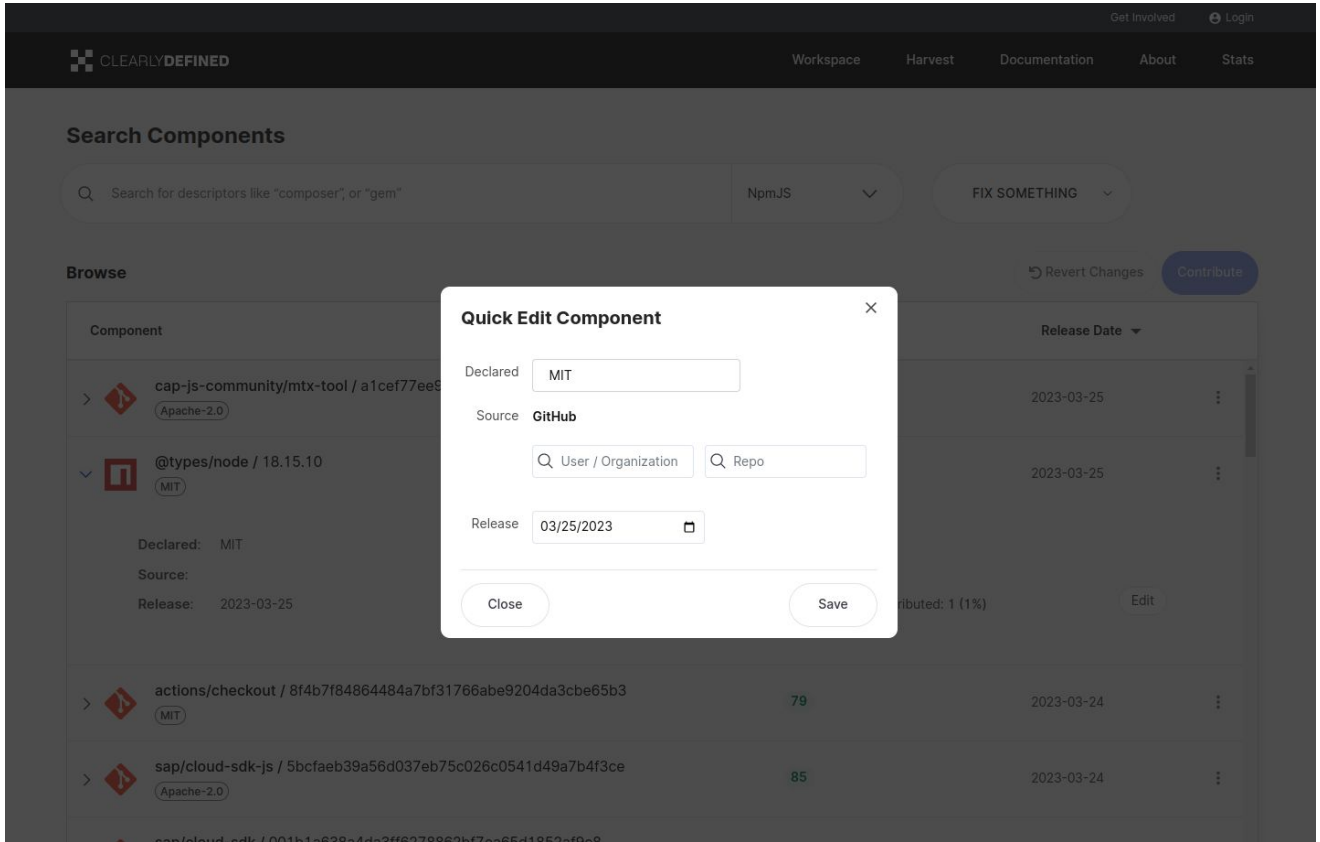
https://api.clearlydefined.io/api-docs/



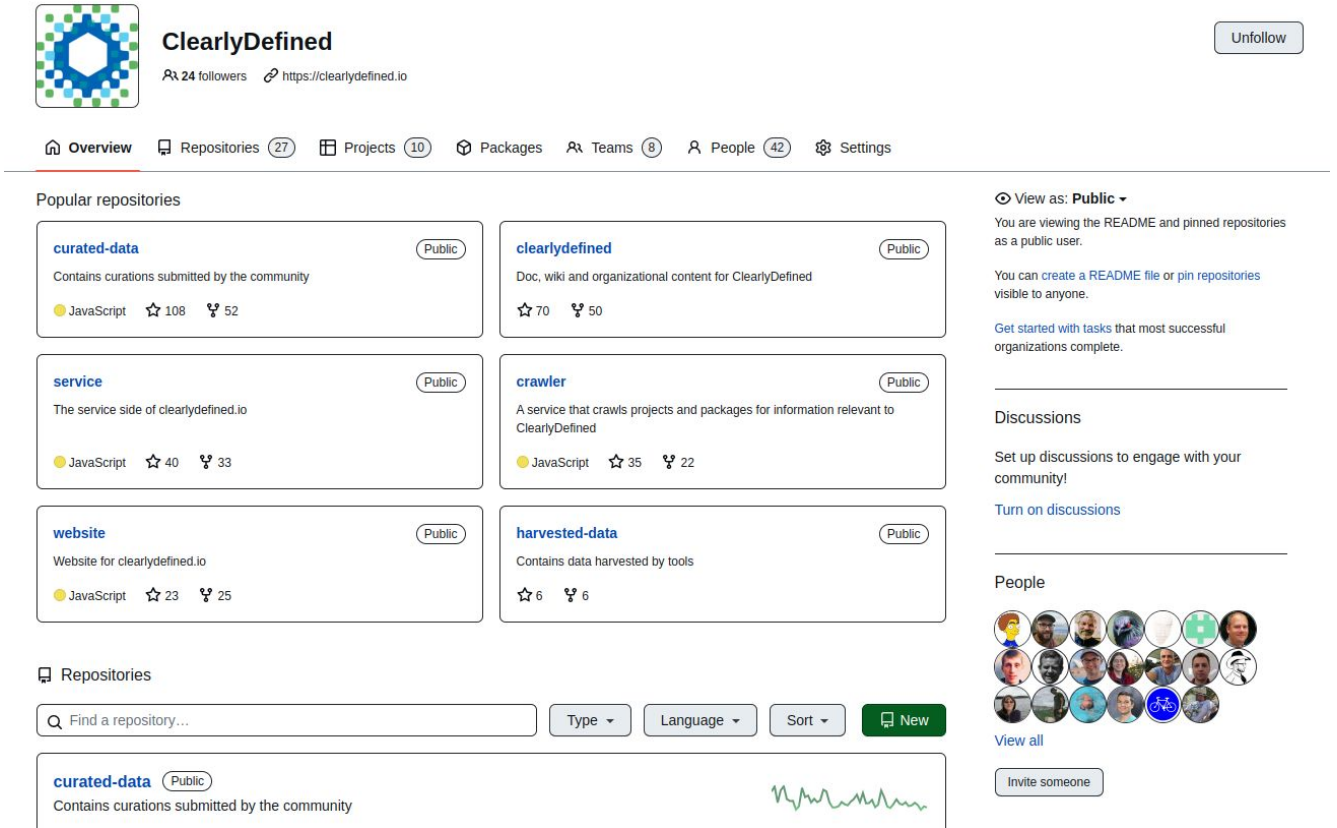
Curate the data


```
"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
 },
```

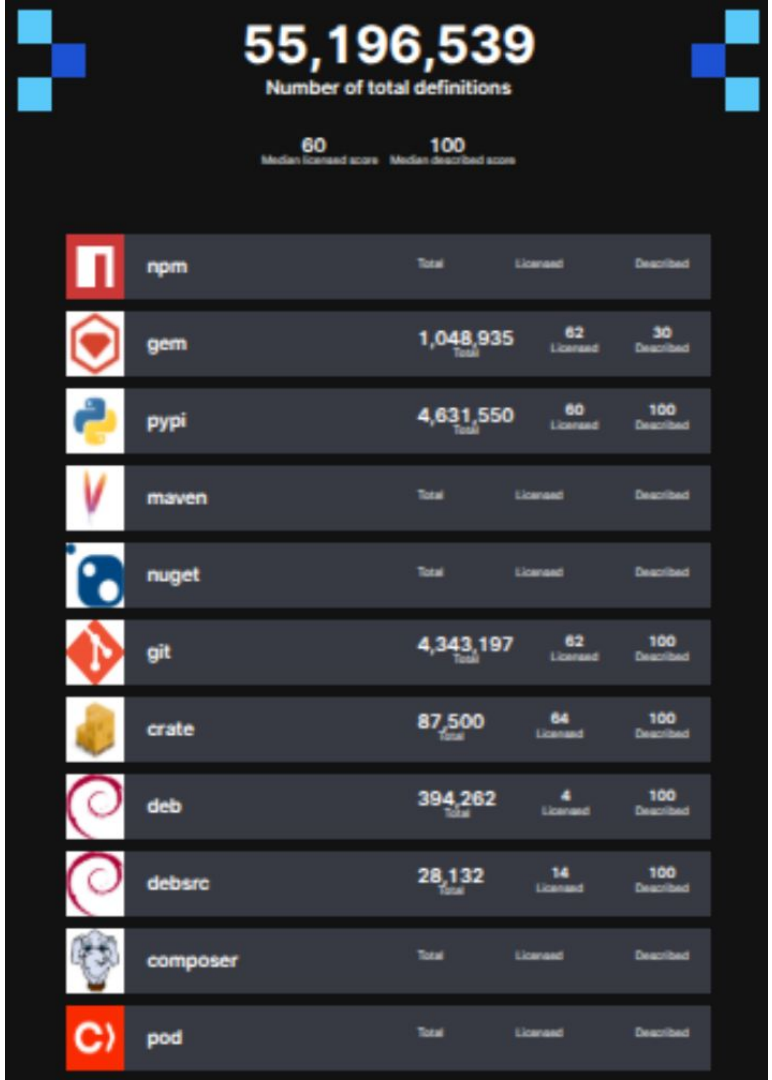
Contribute data



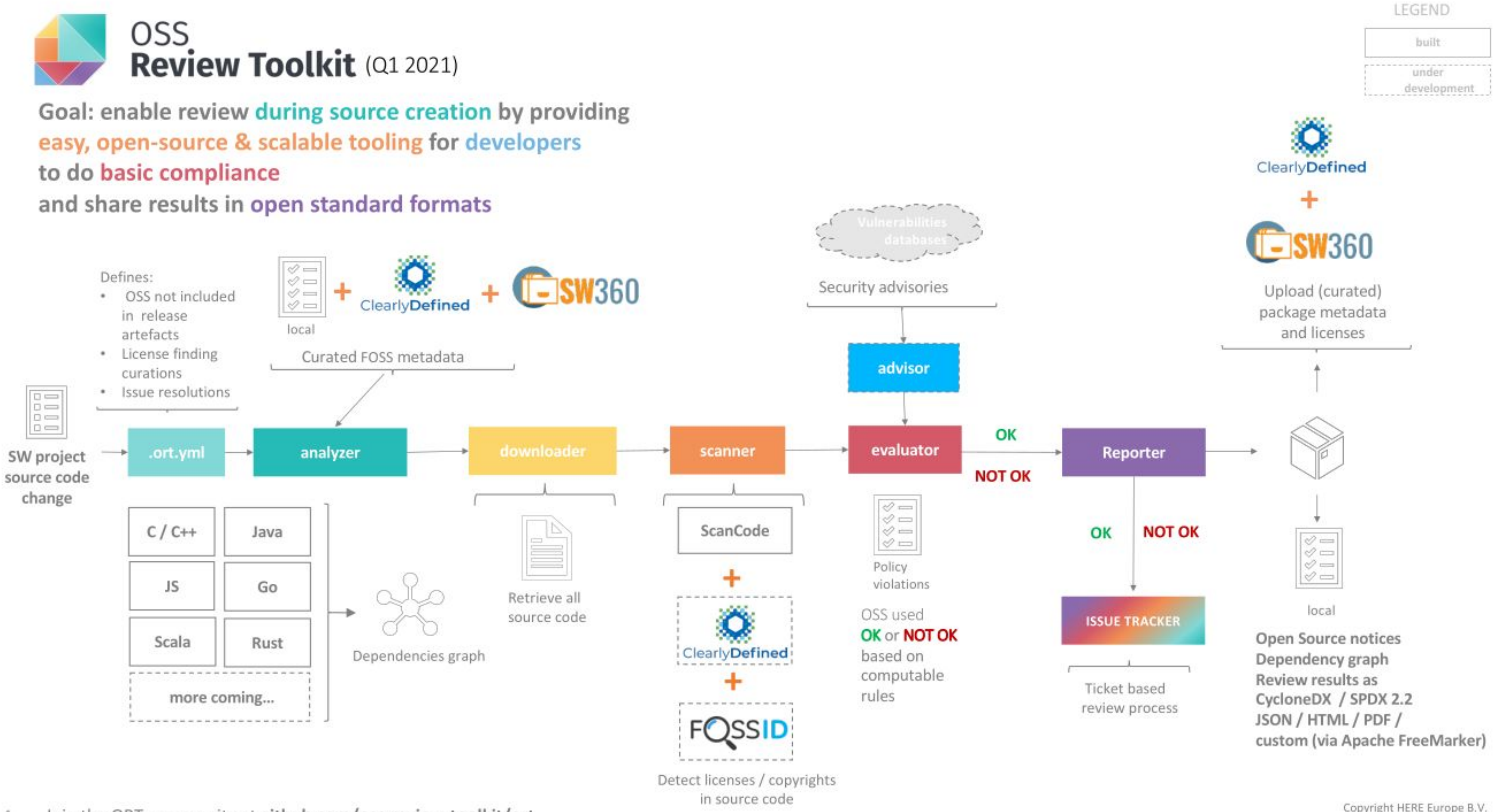
Contribute code



Add a harvest



Adopt best practices



Copyright HERE Europe B.V.
CC-BY--4.0