ClearlyDefined

E. Lynette Rayle (GitHub)

Qing Tomlinson (SAP)

Nick Vidal (Open Source Initiative)

Introduction

Mission

ClearlyDefined’s mission is to crowdsource a global database of licensing metadata for every software component ever published for the benefit of all

Problem

With the move towards SBOMs everywhere for compliance and security reasons, organizations will face great challenges to generate these at scale for each stage on the supply chain, for every build or release.

Plus, multiple organizations will have to fix the same missing or wrongly identified licensing metadata over and over again.

Solution

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

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.

CLEARLYDEFINED

Bringing clarity to Open Source Software licenses.

Use the Data

Curate Data

Contribute Data

Contribute Code

Add a Harvest

Adopt Practices

Use the data

API: definitions, curations, harvest, attachments, notices

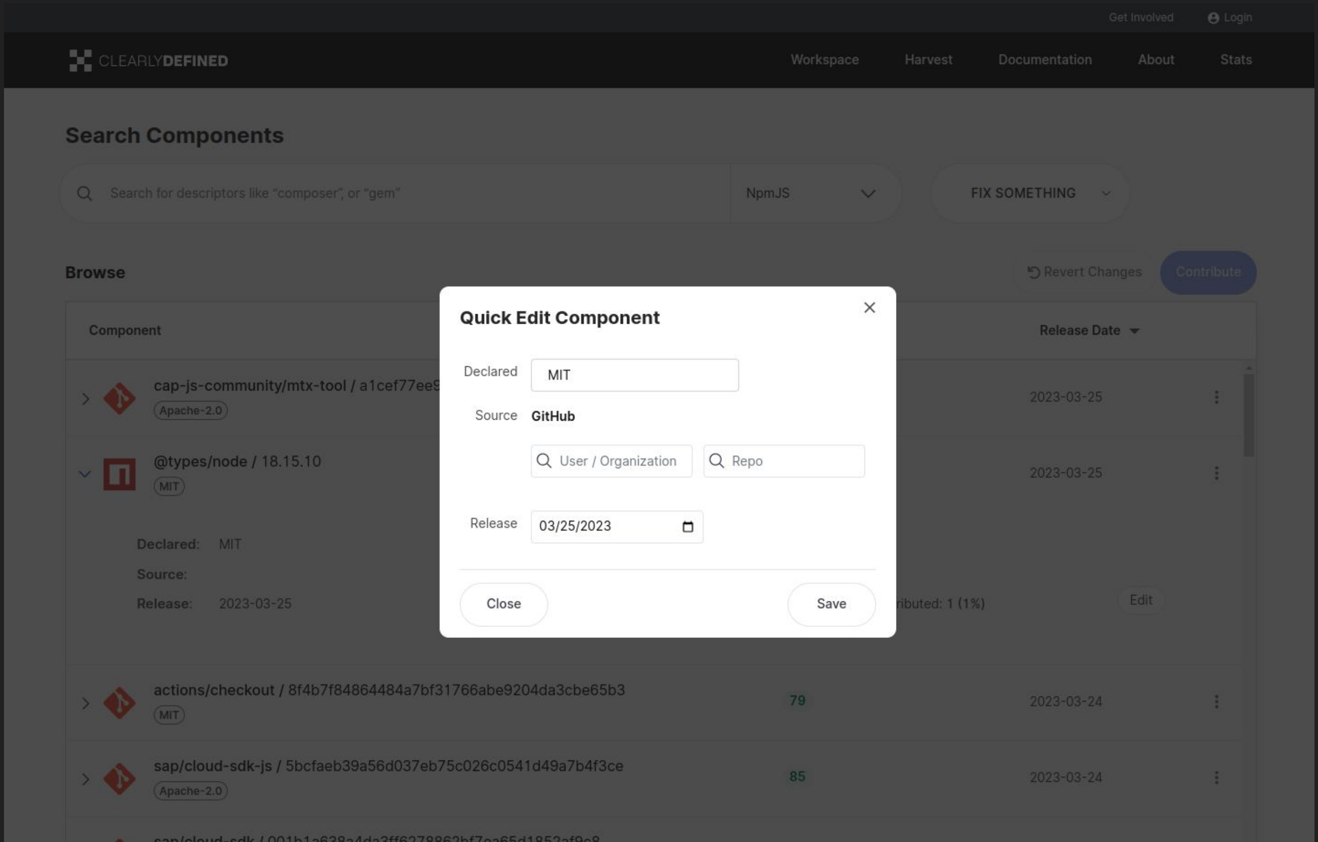
```
curl -X GET
"https://api.clearlydefined.io/definitions/npm/npmjs/-/lodash/4.17.21" -H
"accept: */*"
```

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



Search Components

Search for descriptors like "composer", or "gem"

NpmJS

Browse

Component		Release Date
cap-js-community/mtx-tool / a1cef77ee9 Apache-2.0		2023-03-25
@types/node / 18.15.10 MIT		2023-03-25
actions/checkout / 8f4b7f84864484a7bf31766abe9204da3cbe65b3 MIT	79	2023-03-24
sap/cloud-sdk-js / 5bcfaeb39a56d037eb75c026c0541d49a7b4f3ce Apache-2.0	85	2023-03-24
sap/cloud-sdk / 001b1a638a4da3ff6278862bf7ea65d1852af9c8		

Declared: MIT

Source:

Release: 2023-03-25

ributed: 1 (1%)

Quick Edit Component

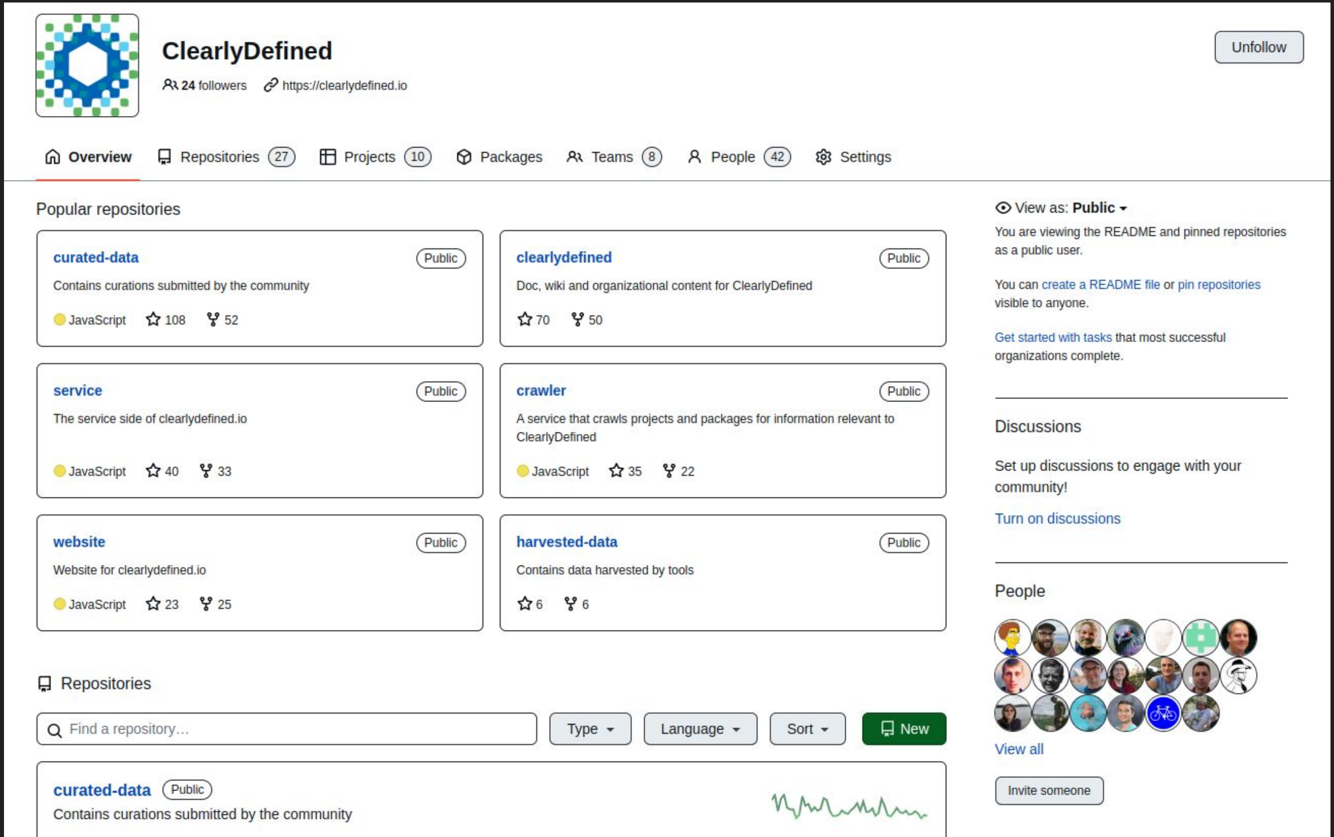
Declared: MIT

Source: GitHub

User / Organization | Repo

Release: 03/25/2023

Contribute code



ClearlyDefined

24 followers https://clearlydefined.io

Unfollow

Overview Repositories 27 Projects 10 Packages Teams 8 People 42 Settings

Popular repositories

curated-data Public
Contains curations submitted by the community
● JavaScript ☆ 108 Forks 52

clearlydefined Public
Doc, wiki and organizational content for ClearlyDefined
☆ 70 Forks 50

service Public
The service side of clearlydefined.io
● JavaScript ☆ 40 Forks 33

crawler Public
A service that crawls projects and packages for information relevant to ClearlyDefined
● JavaScript ☆ 35 Forks 22

website Public
Website for clearlydefined.io
● JavaScript ☆ 23 Forks 25

harvested-data Public
Contains data harvested by tools
☆ 6 Forks 6

Repositories

Find a repository… Type ▾ Language ▾ Sort ▾ New

curated-data Public
Contains curations submitted by the community

View as: **Public** ▾

You are viewing the README and pinned repositories as a public user.

You can create a README file or pin repositories visible to anyone.

Get started with tasks that most successful organizations complete.

Discussions

Set up discussions to engage with your community!

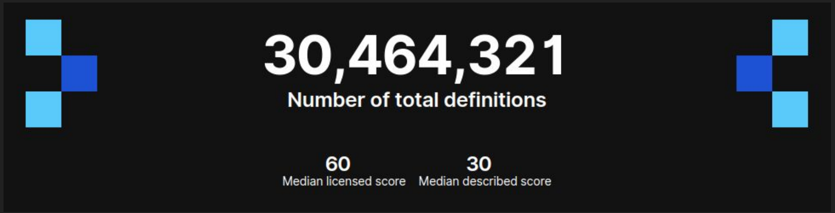
Turn on discussions

People

View all

Invite someone

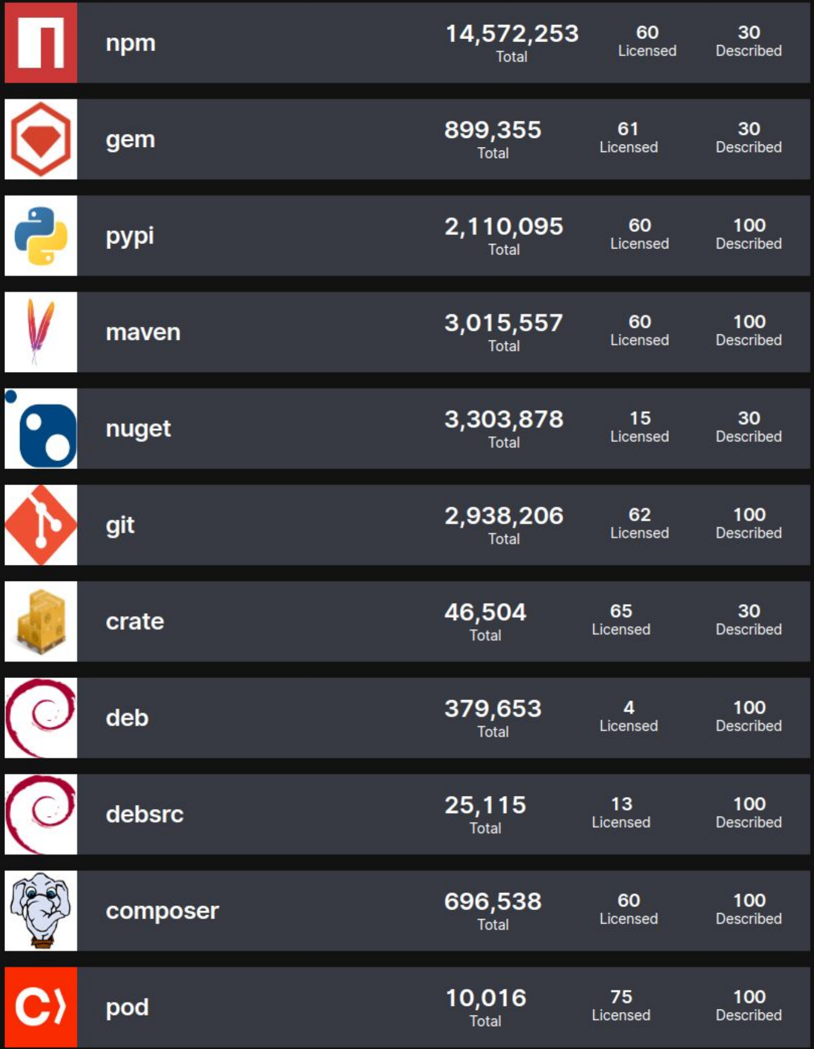
Add a harvest



30,464,321
Number of total definitions

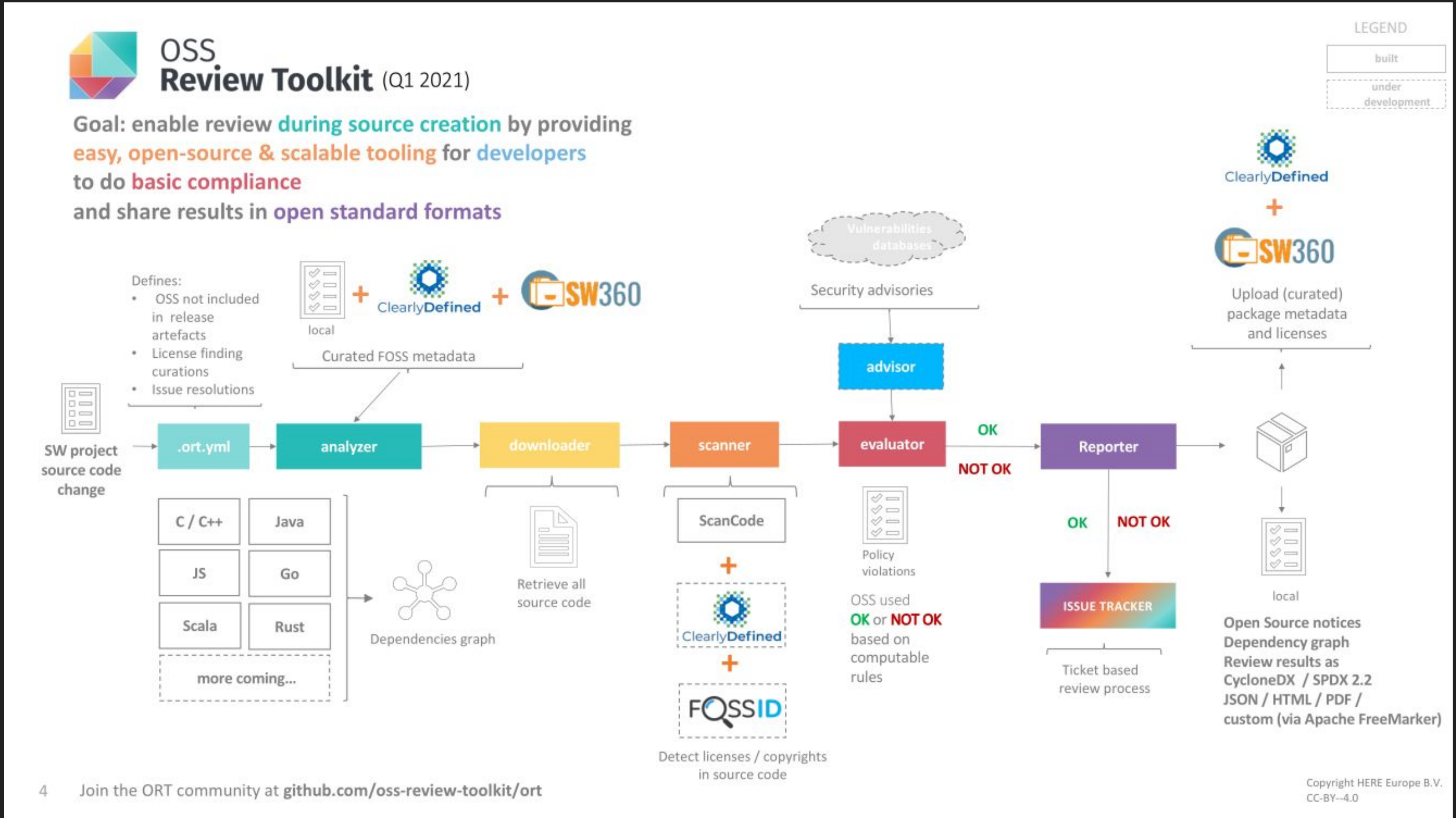
60
Median licensed score

30
Median described score



npm	14,572,253 Total	60 Licensed	30 Described
gem	899,355 Total	61 Licensed	30 Described
pypi	2,110,095 Total	60 Licensed	100 Described
maven	3,015,557 Total	60 Licensed	100 Described
nuget	3,303,878 Total	15 Licensed	30 Described
git	2,938,206 Total	62 Licensed	100 Described
crate	46,504 Total	65 Licensed	30 Described
deb	379,653 Total	4 Licensed	100 Described
debsrc	25,115 Total	13 Licensed	100 Described
composer	696,538 Total	60 Licensed	100 Described
pod	10,016 Total	75 Licensed	100 Described

Adopt best practices



OSS
Review Toolkit (Q1 2021)

Goal: enable review during source creation by providing easy, open-source & scalable tooling for developers to do basic compliance and share results in open standard formats

4 Join the ORT community at **github.com/oss-review-toolkit/ort**

New developments

LicenseRef

- Beyond SPDX: expanding licenses identified by ClearlyDefined

Conda

- Better identifying conda packages with ClearlyDefined

GUAC

- GUAC adopts license metadata from ClearlyDefined
  Announcing GUAC v0.8.0 Enhancements

Switch to Lynette

Case Studies

A Developer's Look at ClearlyDefined

and how GitHub is using it and why

Why we’re using ClearlyDefined…

ClearlyDefined houses **business critical data** for licenses and attributions. We want to support the mission of making ClearlyDefined **THE** source of truth.

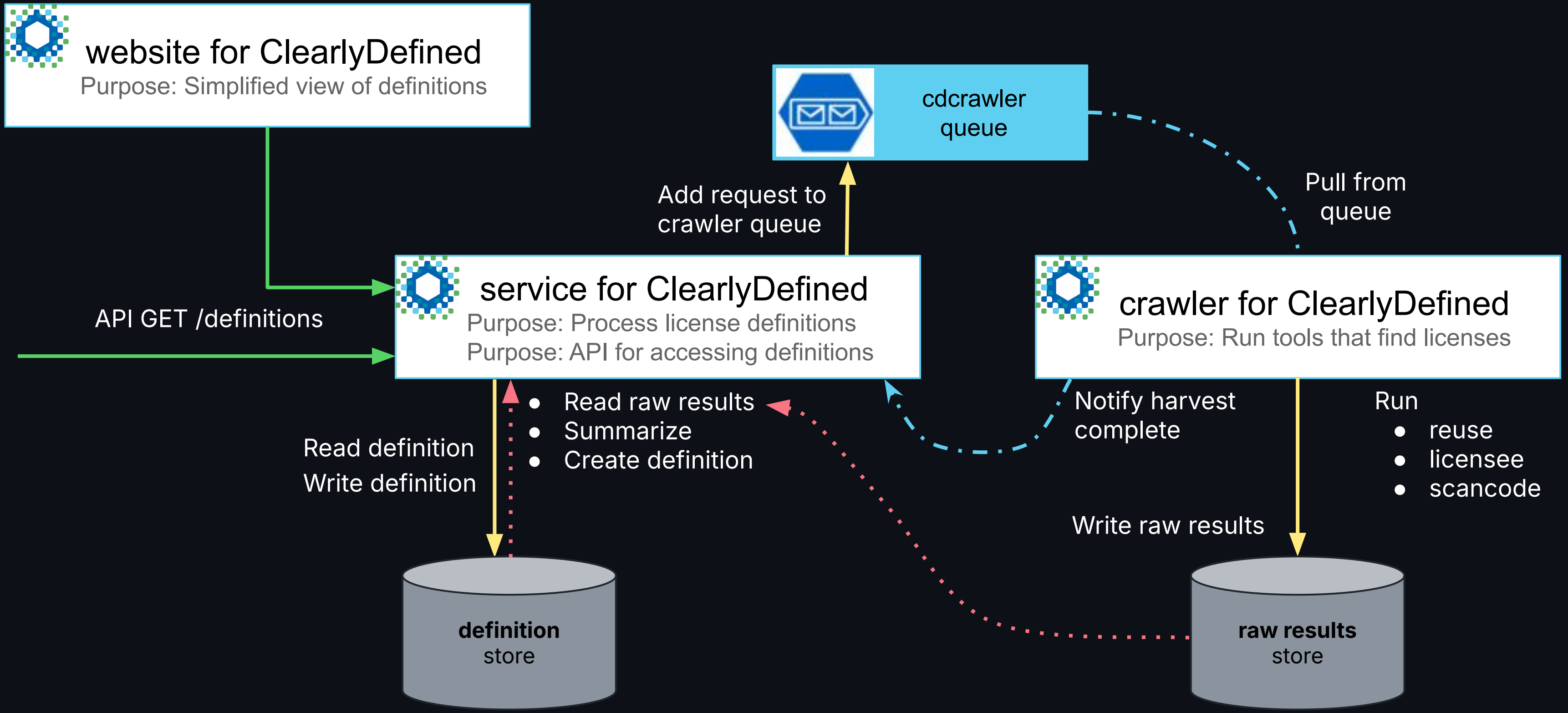
Impact of ClearlyDefined at GitHub

GitHub added 17.5 million package licenses

- sourced from ClearlyDefined to our database,
- expanding the license coverage for packages that appear in
  - dependency graph
  - dependency insights
  - dependency review
  - repository’s software bill of materials (SBOM)

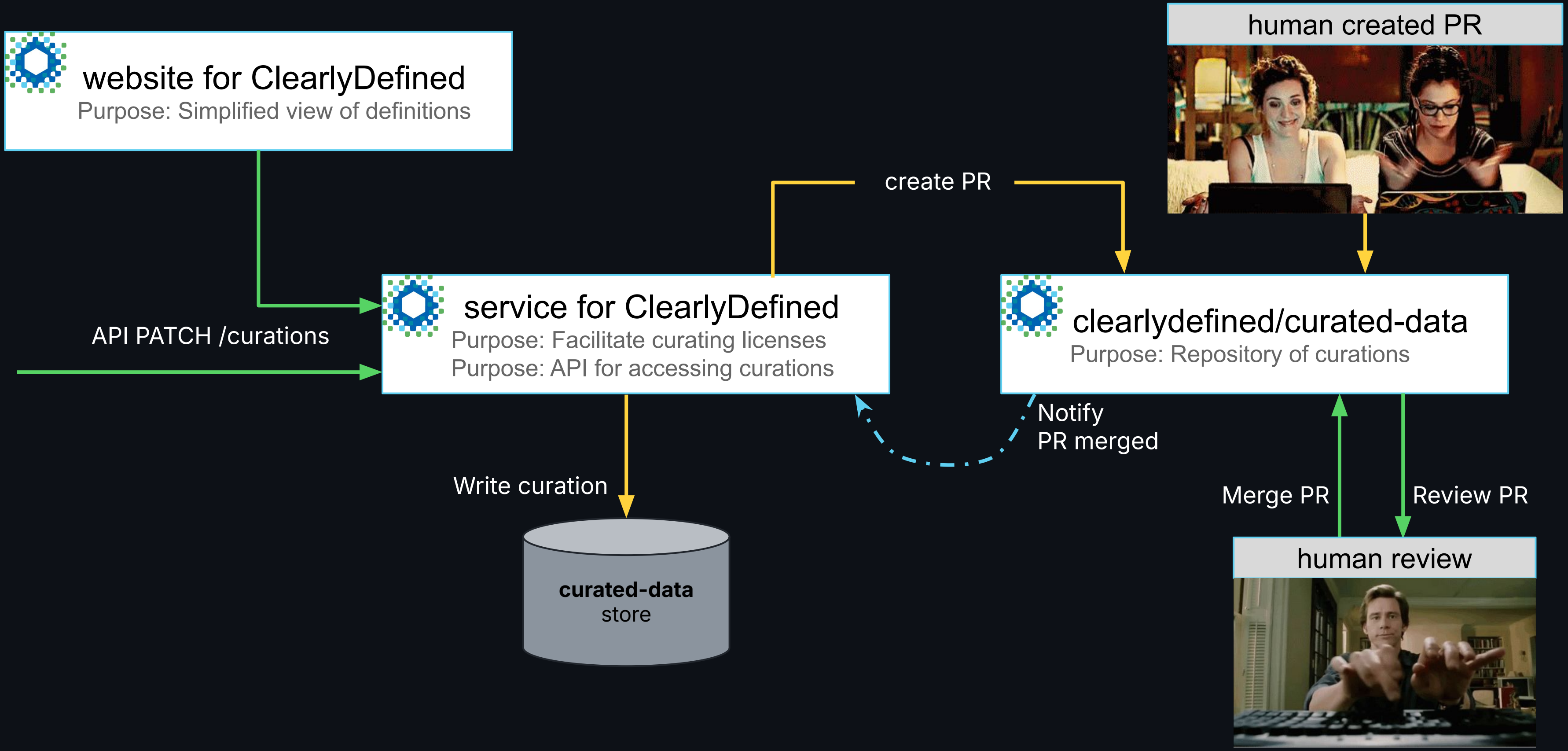
Crawler == Harvester

Overview of ClearlyDefined Harvesting Process


website for ClearlyDefined
Purpose: Simplified view of definitions
cdcrawler queue
Add request to crawler queue
Pull from queue
API GET /definitions
service for ClearlyDefined
Purpose: Process license definitions
Purpose: API for accessing definitions
crawler for ClearlyDefined
Purpose: Run tools that find licenses
Read raw results
Summarize
Create definition
Notify harvest complete
Run
reuse
licensee
scancode
Read definition
Write definition
Write raw results
definition store
raw results store

Curation == Human

Overview of ClearlyDefined Curation Process



website for ClearlyDefined

Purpose: Simplified view of definitions

API PATCH /curations

service for ClearlyDefined

Purpose: Facilitate curating licenses

Purpose: API for accessing curations

create PR

human created PR

clearlydefined/curated-data

Purpose: Repository of curations

Notify

PR merged

Write curation

curated-data

store

Merge PR

Review PR

human review

Clearly Defined Change Notifications

changes-notifications
Azure blob storage
(public read access)

```
|- changes
|  |- index
|  |- 2024-02-29-22
|  |- 2024-02-29-23
|  |- 2024-03-01-00
|  |- ...
|- gem
|  |- rubygems
|  |  |- -
|  |  |  |- rspec
|  |  |  |  |- revisions
.  |  |  |  |  |- 3.13.0.json
|- ...
```

index - lists all changeset IDs past and present

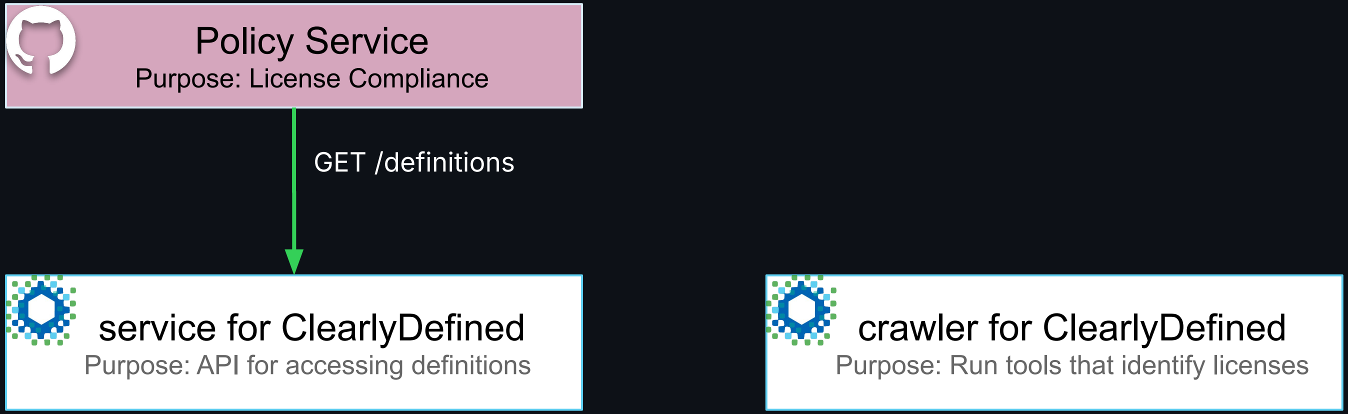
changeset - lists coordinates of changes named for the date and hour when the changeset was created

example: 2024-02-29-22 is Feb 29, 2024 at 2200 hours

path based on coordinates

definition - package version as a json file holding the definition

GitHub & ClearlyDefined



Policy Service

Purpose: License Compliance

GET /definitions

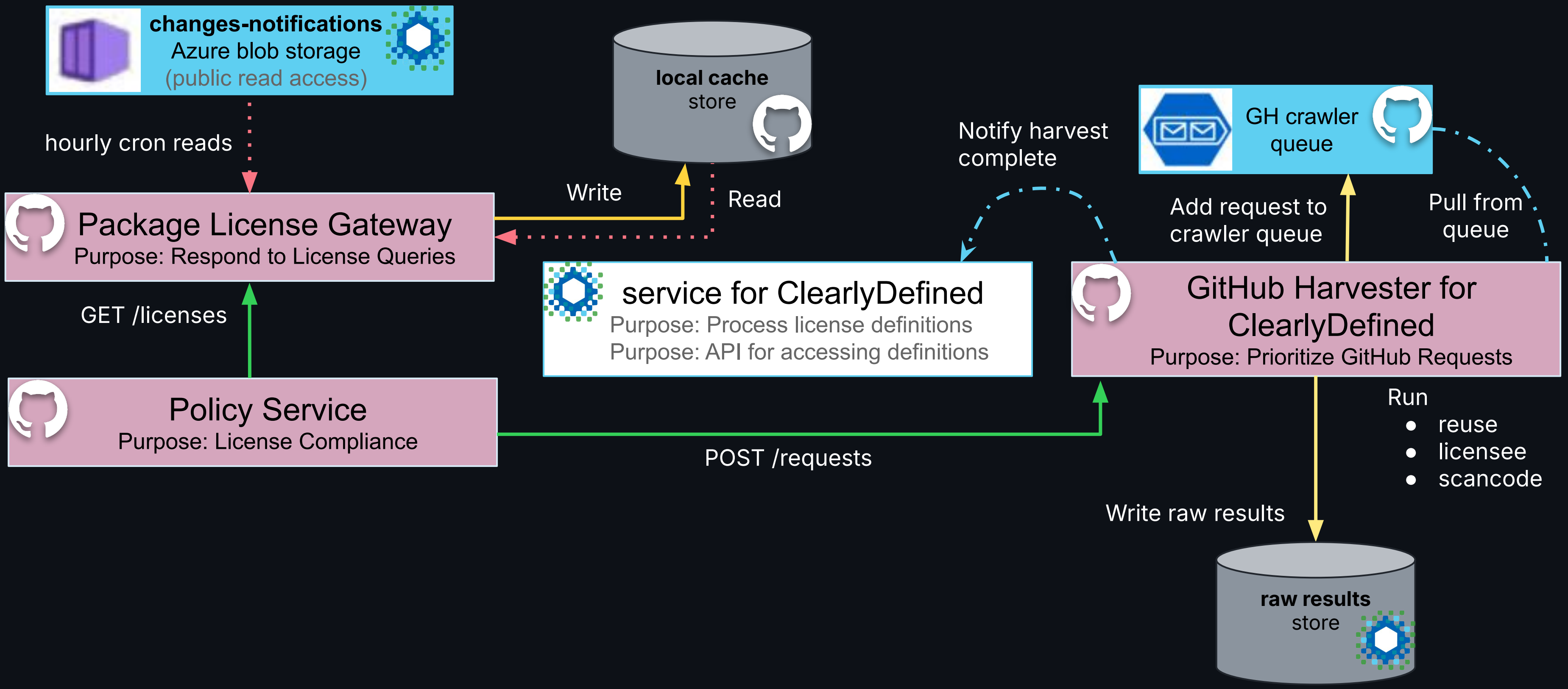
service for ClearlyDefined

Purpose: API for accessing definitions

crawler for ClearlyDefined

Purpose: Run tools that identify licenses

Handling High Volume of Requests



changes-notifications
Azure blob storage
(public read access)

hourly cron reads

local cache
store

Write

Read

Package License Gateway
Purpose: Respond to License Queries

GET /licenses

Policy Service
Purpose: License Compliance

POST /requests

service for ClearlyDefined
Purpose: Process license definitions
Purpose: API for accessing definitions

Notify harvest complete

GH crawler queue

Add request to crawler queue

Pull from queue

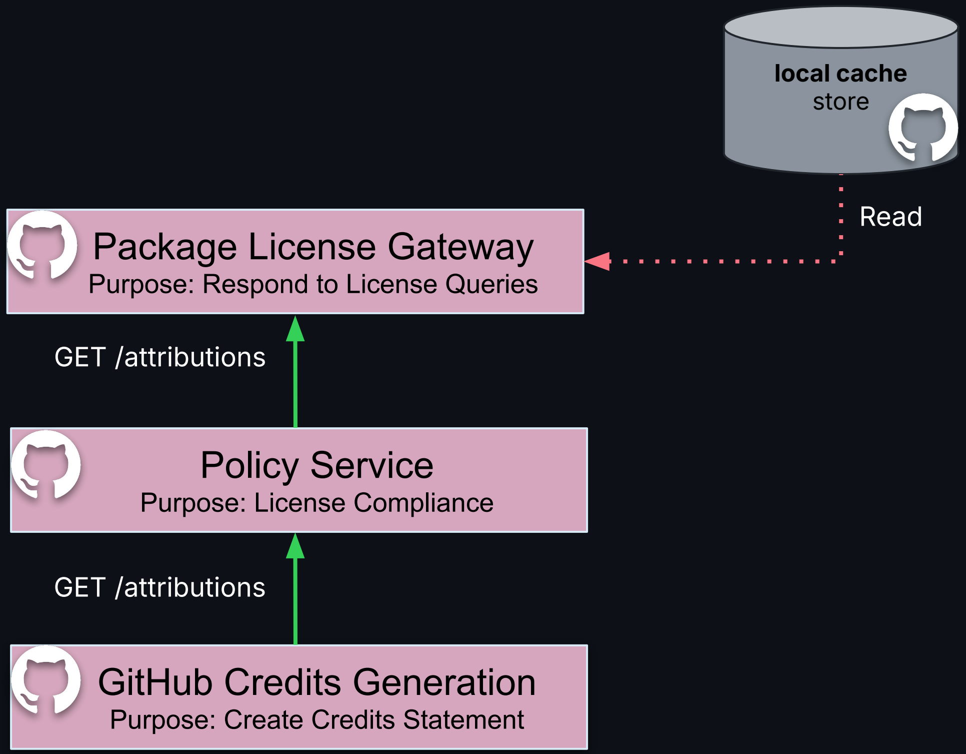
GitHub Harvester for ClearlyDefined
Purpose: Prioritize GitHub Requests

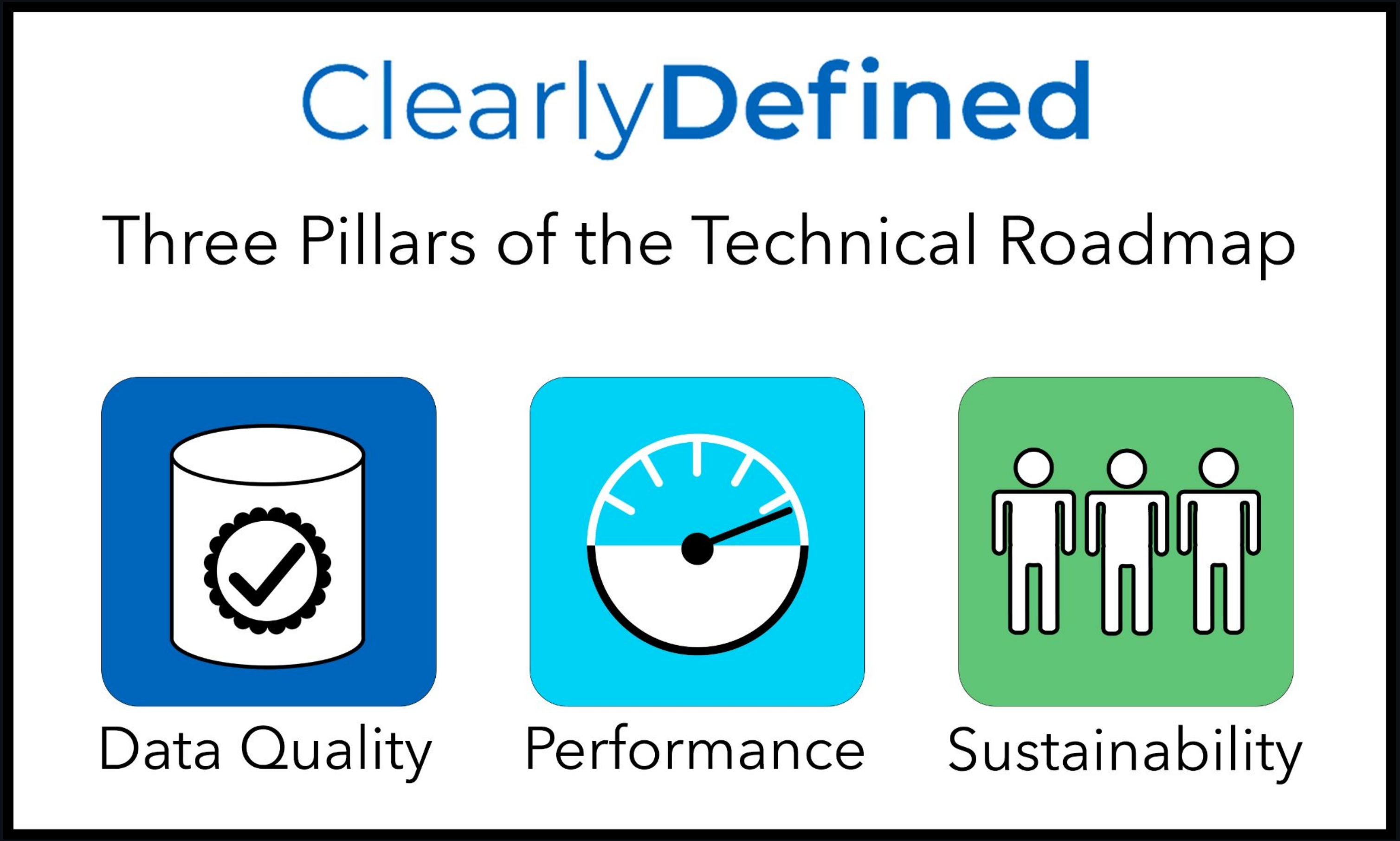
Run
- reuse
- licensee
- scancode

Write raw results

raw results
store

Complying with Attribution Requirements


local cache
store
Read
Package License Gateway
Purpose: Respond to License Queries
GET /attributions
Policy Service
Purpose: License Compliance
GET /attributions
GitHub Credits Generation
Purpose: Create Credits Statement



ClearlyDefined

Three Pillars of the Technical Roadmap

Data Quality

Performance

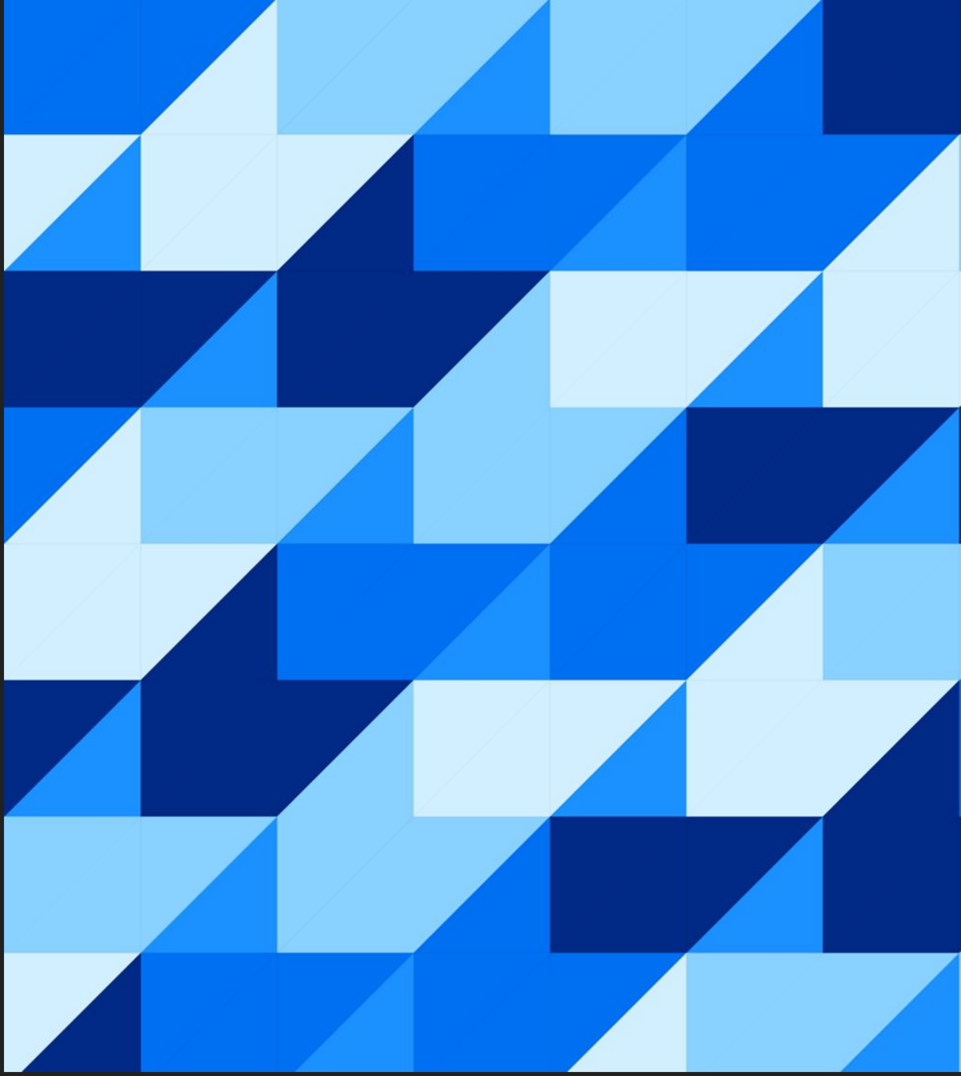
Sustainability

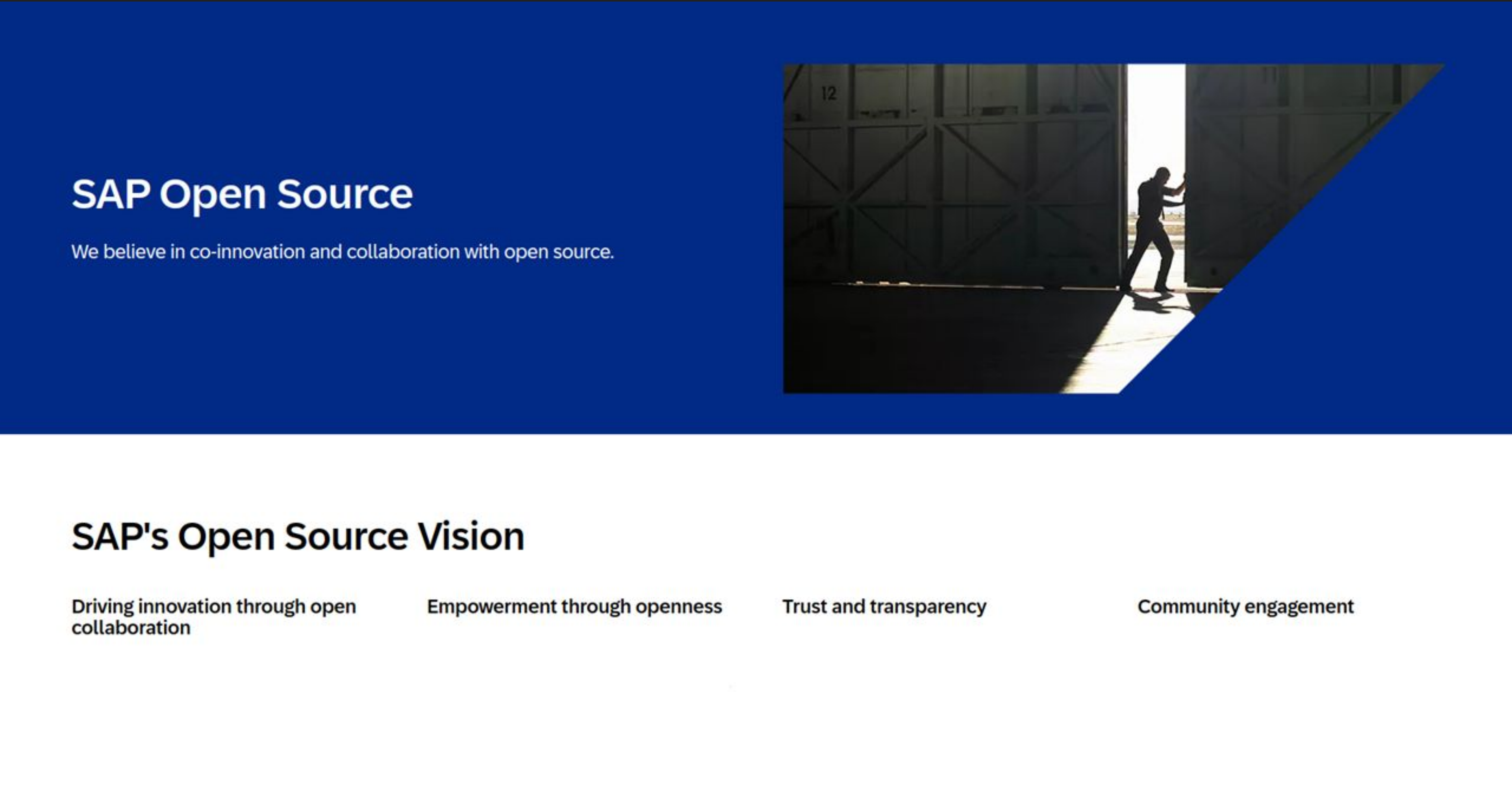
How we setup a harvester at GitHub

- Used kubernetes config in crawler.yaml (in clearlydefined/crawler repo)
  - uses official crawler image in Docker Hub
    (may be moving to GitHub container registry)
  - lists configurations to be set
- Requires a token to write directly to ClearlyDefined's raw result store
- Requires significant hardware
  - Ex. Azure P3V2 (4 Virtual CPU's, 14GB Ram)
    (values for ClearyDefined's production crawler)

Switch to Qing

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

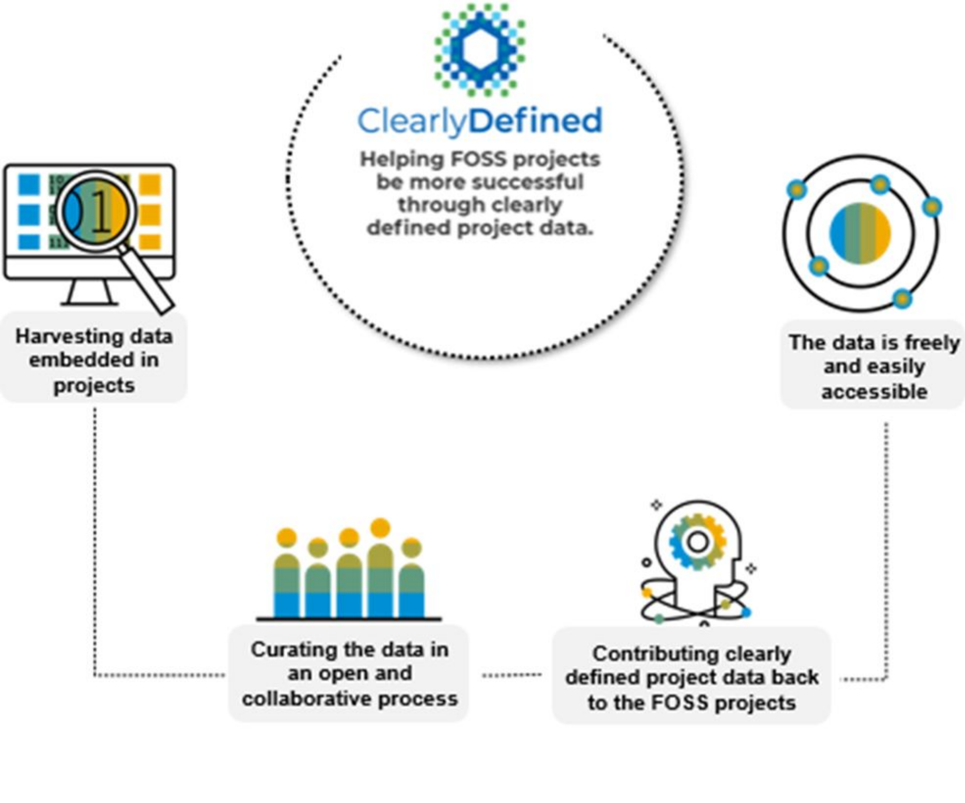
Empowerment through openness

Trust and transparency

Community engagement

Why ClearlyDefined?

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



ClearlyDefined

Helping FOSS projects be more successful through clearly defined project data.

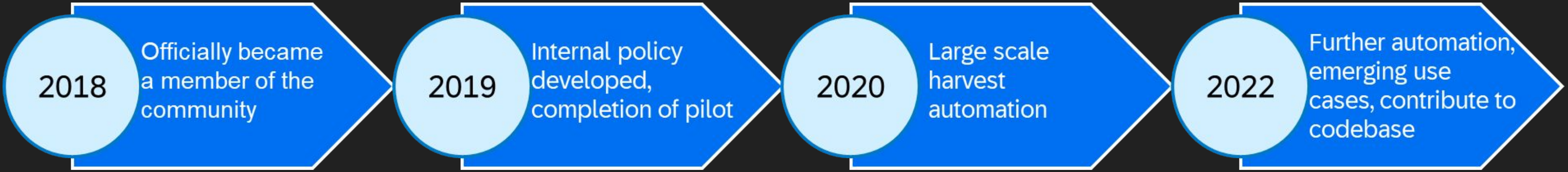
Harvesting data embedded in projects

Curating the data in an open and collaborative process

Contributing clearly defined project data back to the FOSS projects

The data is freely and easily accessible

ClearlyDefined Adoption – Key Milestones



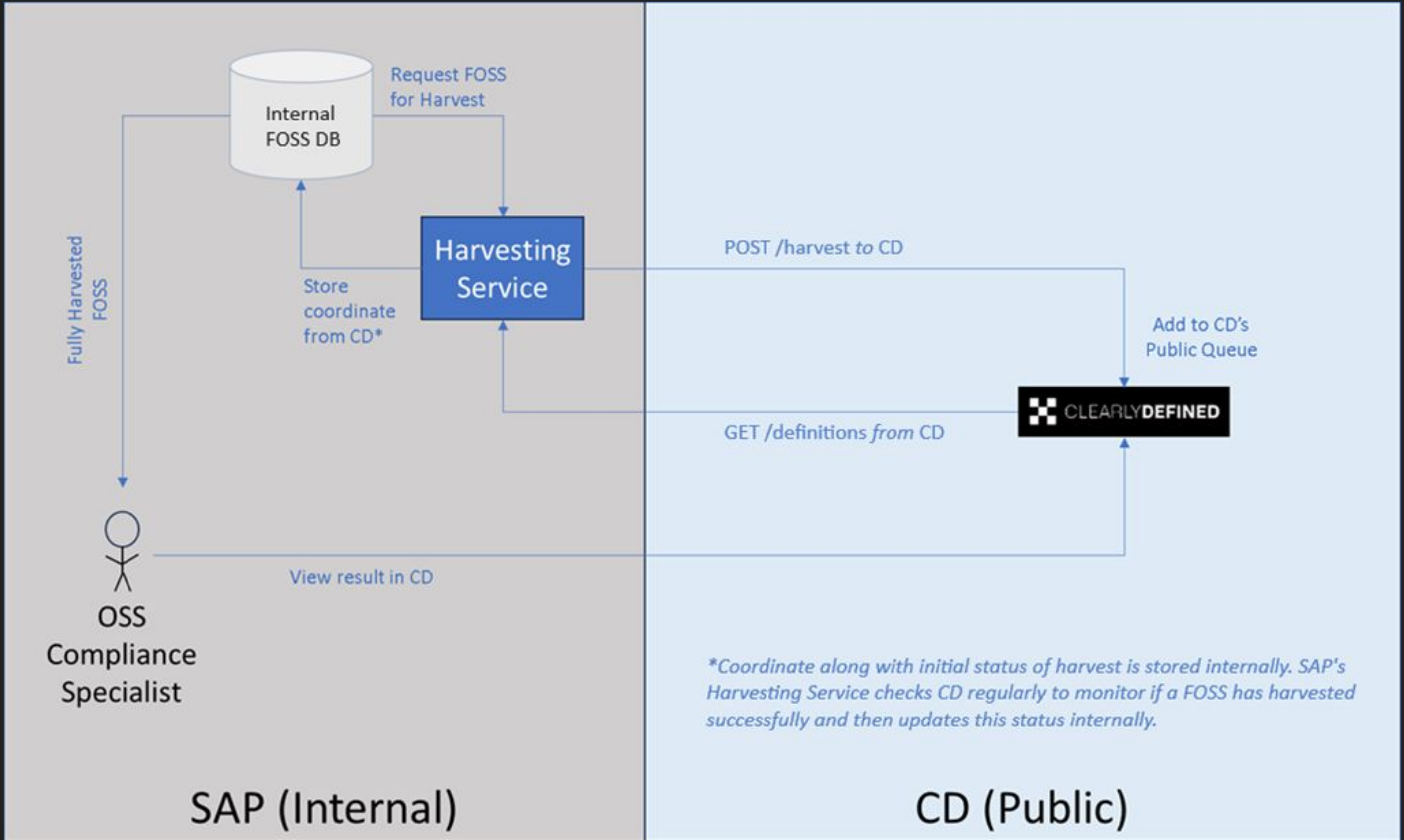
2018
Officially became a member of the community

2019
Internal policy developed, completion of pilot

2020
Large scale harvest automation

2022
Further automation, emerging use cases, contribute to codebase

Overview of Automatic Harvesting



**Coordinate along with initial status of harvest is stored internally. SAP's Harvesting Service checks CD regularly to monitor if a FOSS has harvested successfully and then updates this status internally.*

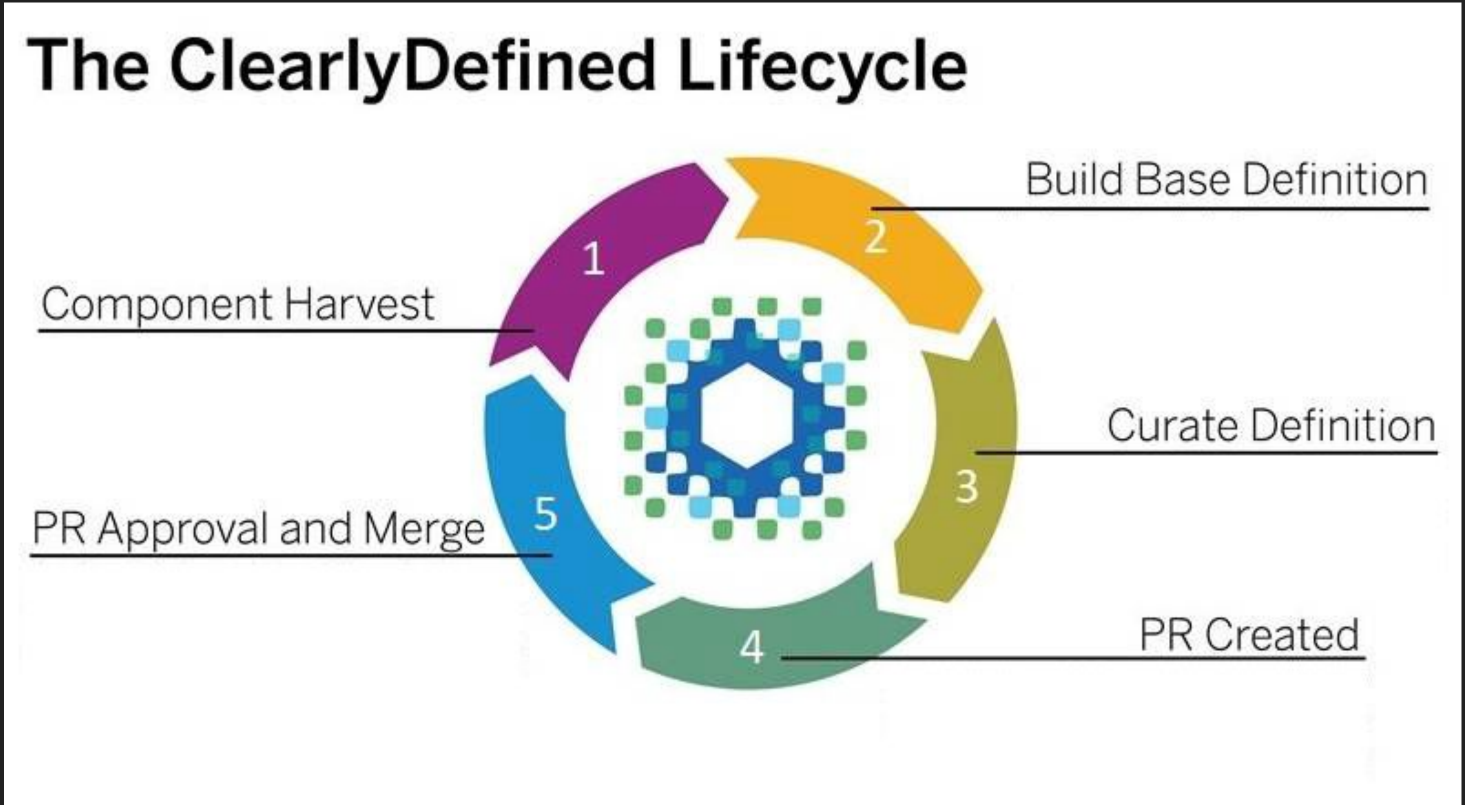
ClearlyDefined Benefits

Less time scanning, more time focusing on compliance

Better data quality

Community centered

Giving Back – Contributions to Data



The ClearlyDefined Lifecycle

1 Component Harvest

2 Build Base Definition

3 Curate Definition

4 PR Created

5 PR Approval and Merge

Giving Back – Contributions to Code

Harvest
Gradle Plugin

Optimizations

Analyses and
Improvements

Collaborations

Key Takeaways

30-50% Improvement

Community

Switch to Nick

Governance

Governance

ClearlyDefined Governance:

- Charter

Governing Board:

- the Executive Director of the Open Source Initiative;
- the Steering Committee Chair; and
- the Outreach Committee Chair.

Governance

Steering Committee:

- E. Lynette Rayle (GitHub) - Chair
- Qing Tomlinson (SAP)
- Jeff Mendoza (GUAC / Kusari)

Governance

Outreach Committee:

- Nick Vidal (OSI) - Chair
- Jeff Luszcz (GitHub)
- Brian Duran (SAP)
- Alyssa Wright (Bloomberg)

Conclusion

How to get involved...

- Weekly dev meetings
- https://docs.clearlydefined.io/docs/community/meetings
- Hang out in Discord
- Paired programming
- Understand the inner working of ClearlyDefined
- Influence the priorities of development
- Help sustain and keep ClearlyDefined strong
- Set up a harvester on your hardware

Open Collaboration

OSS Review Toolkit
OPENCHAIN
SBOM
SW360
SPDX
CycloneDX
ScanCode
fossology
TODO
GUAC
CSAF

Thank you

Join us:

https://clearlydefined.io/

ClearlyDefined