



ClearlyDefined



GitHub

E. Lynette Rayle

Senior Developer



**open source
initiative®**

Nick Vidal

Community Manager

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.



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

Clearly Defined

Workspace Harvest Documentation About Stats

Get Involved Login

Search Components

Search for descriptors like "composer", or "gem" NpmJS FIX SOMETHING

Browse

Revert Changes Contribute

Quick Edit Component

Declared: MIT

Source: GitHub

User / Organization Repo


Release: 03/25/2023

Close Save

Component	Release Date
cap-js-community/mtx-tool / a1cef77ee5 (Apache-2.0)	2023-03-25
@types/node / 18.15.10 (MIT)	2023-03-25
Declared: MIT Source: Release: 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 / 001b1e639a4da2ff6278962b7ca65d1852af0e8	



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 🍴 52

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

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

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


website Public
Website for clearlydefined.io
JavaScript ☆ 23 🍴 25

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

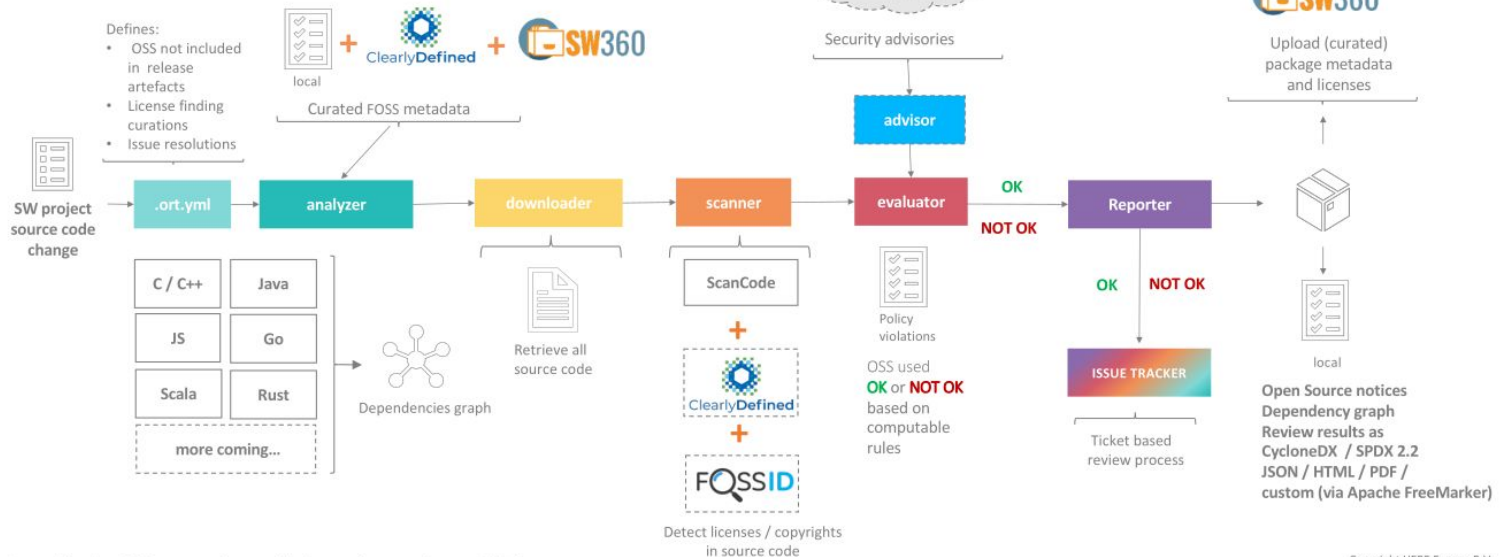


Adopt best practices



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**



Switch to Lynette



A Developer's Look at ClearlyDefined

and a little about 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.



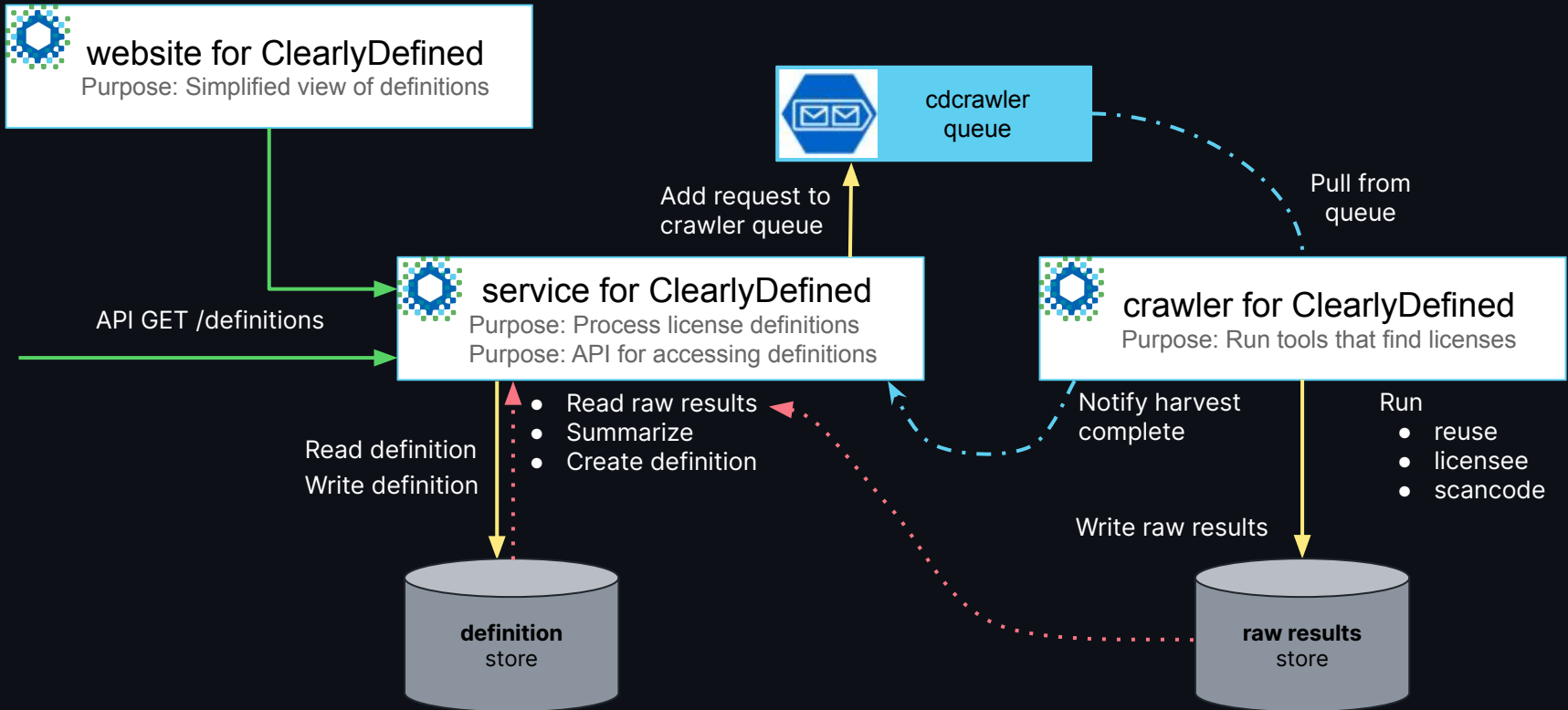
Impact of ClearlyDefined at GitHub

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)

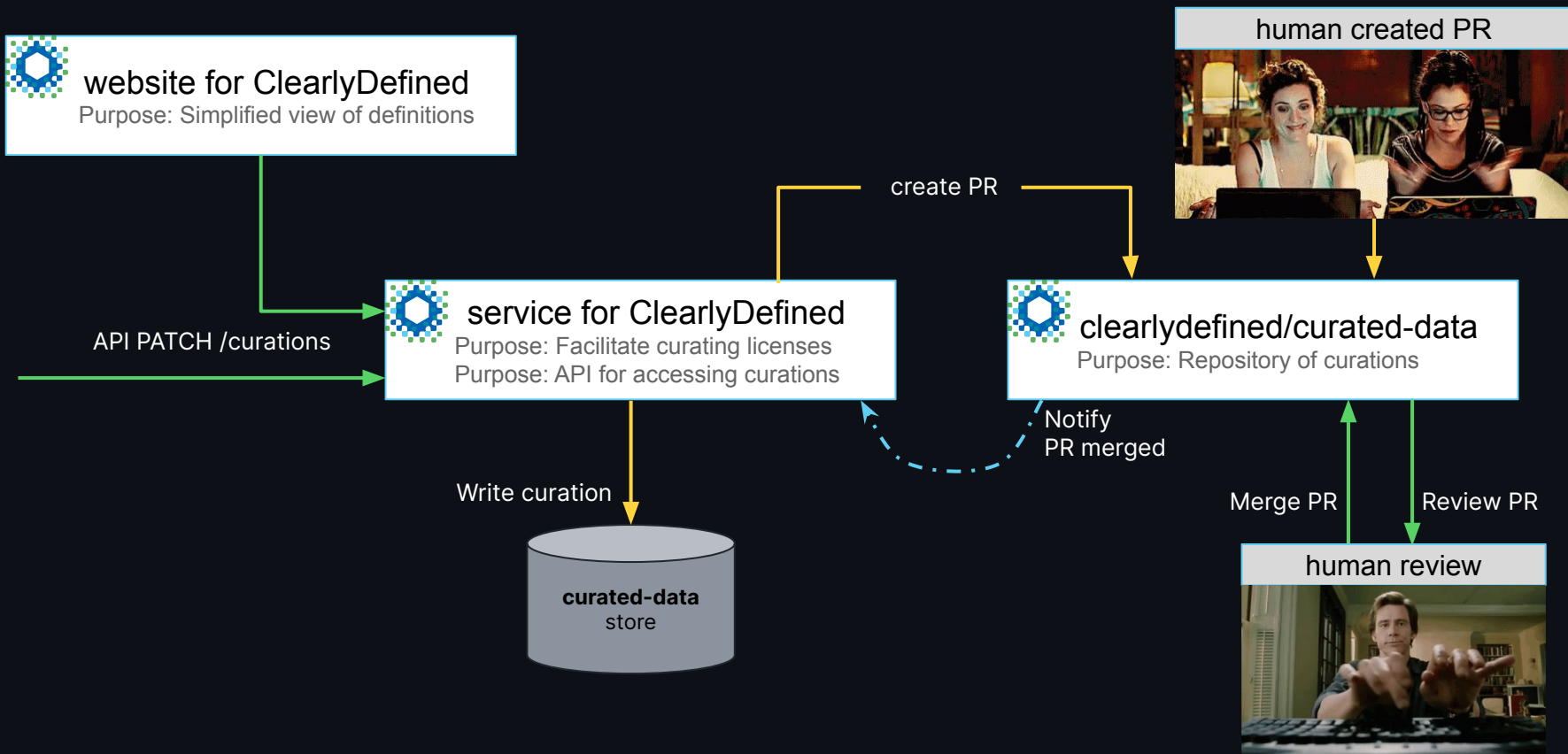
Crawler == Harvester

Overview of ClearlyDefined Harvesting Process

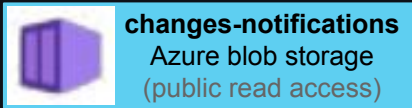


Curation == Human

Overview of ClearlyDefined Curation Process

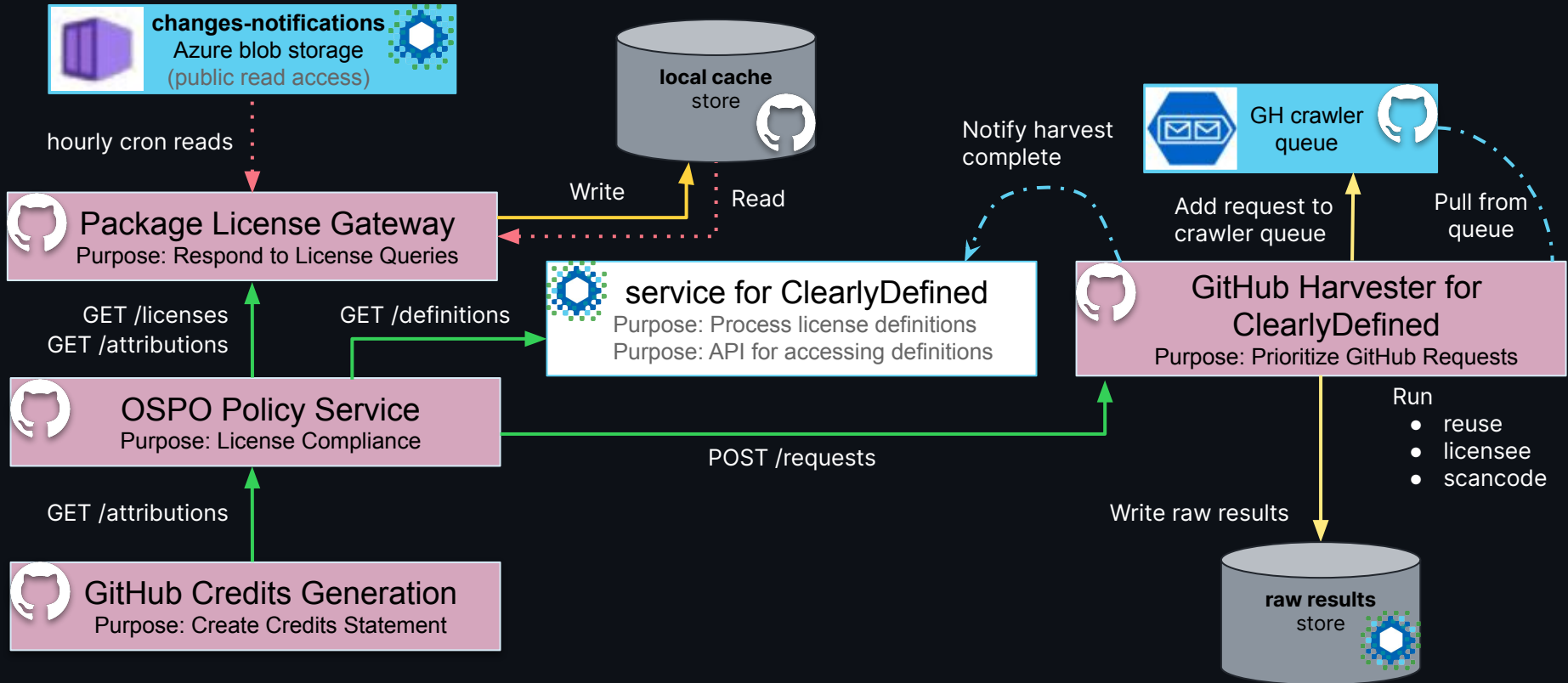


Clearly Defined Change Notifications



- 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

High Level Overview of GitHub & ClearlyDefined





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 ClearlyDefined's production crawler)

What's happening...

in ClearlyDefined Development



Ongoing Maintenance

- Critical updates to keep tools that identify licenses at the latest revision (GitHub, SAP)
- Address outdated dependencies (GitHub, SAP, Microsoft)
- Automated integration tests (SAP)
- Bug fixes (SAP, GitHub, Microsoft)
- Documentation (SAP, GitHub, OSI, Microsoft)



Working on Now

- Add support for latest scancode & licensee (GitHub, SAP)
- Addition of Conda as a source (CodeThink, SAP)
- Move deploy to GitHub actions (GitHub)
- Use semantic versioning for releases (GitHub)
- Add GitHub action to regularly update licenses from SPDX (GitHub)
- Bug fixes leading to resolution of some licenses classified as NOASSERTION (SAP)
- Move production crawler to ClearlyDefined Azure space



What next? The wish list...

- License Data Quality
- Reporting of LicenseRef
- Curation of Attributions
- Improve Throughput
- Higher Rate Limits
- UI Usability Improvements
- Better Monitoring
- Support other entities setting up Harvesters on their hardware



How to get involved...

- Weekly dev meetings
- Monthly 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

Switch to Nick

Conclusion



Conclusion

- Open Development
- Open Data
- Open Governance
- Open Collaboration



Open Development

- A clear development roadmap with identifiable points of contribution
- Enhance the documentation to make use of ClearlyDefined
- Make it easier to set up crawlers (local harvesters)



Open Data

ClearlyDefined is working to provide open access to more data, including raw scan results, regular data exports, and metadata from releases.



Open Governance

ClearlyDefined is working to establish a clear and open governance structure to become more welcoming towards contributors.

<https://www.linkedin.com/pulse/whats-open-governance-drafting-charter-source-project-nick-vidal/>



Open Collaboration



OSS
Review Toolkit

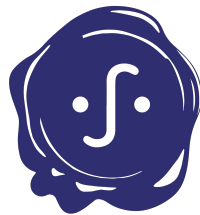
OPENCHAIN



ScanCode



GUAC





Thank you

Join us:

<https://clearlydefined.io/>



ClearlyDefined

