Clearly Defined

E. Lynette Rayle
Senior Developer

Nick Vidal
Community Manager
Introduction
ClearlyDefined’s mission is to crowdsourced 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.
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.
Clearly Defined

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

30,464,321

Number of total definitions

Median licensed score: 60
Median described score: 30

Add a harvest
Adopt best practices
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

- **website for ClearlyDefined**
  - Purpose: Simplified view of definitions

- **API GET /definitions**
- **service for ClearlyDefined**
  - Purpose: Process license definitions
  - Purpose: API for accessing definitions
  - Activities:
    - Read raw results
    - Summarize
    - Create definition

- **cdcrawler queue**
  - Add request to crawler queue

- **crawler for ClearlyDefined**
  - Purpose: Run tools that find licenses
  - Activities:
    - Notify harvest complete
    - Run:
      - reuse
      - licensee
      - scanode

- **raw results store**
- **definition store**
  - Read definition
  - Write definition
  - Write raw results
Curation == Human
Overview of ClearlyDefined Curation Process

website for ClearlyDefined
Purpose: Simplified view of definitions

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

clearlydefined/curated-data
Purpose: Repository of curations

curated-data store

API PATCH /curations

create PR

Notify
PR merged

human review

Merge PR

Review PR

human created PR
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
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)
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 scanocode & 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
Thank you

Join us:
https://clearlydefined.io/

ClearlyDefined