AboutCode

What's new in AboutCode: ScanCode, MatchCode, VulnerableCode, and beyond for ORT users

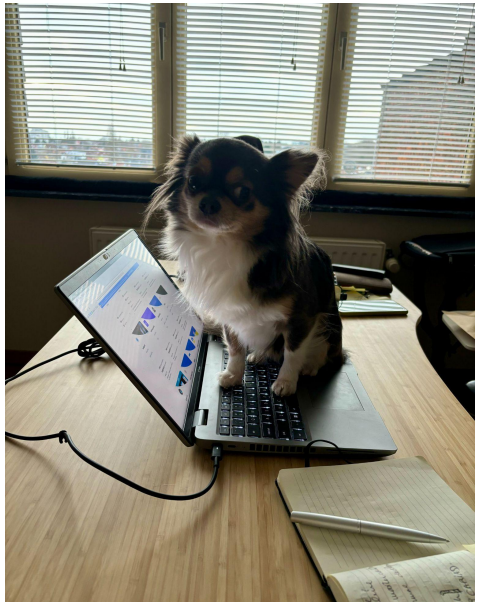
Agenda

AboutCode

- About me, AboutCode
- Software Composition Analysis
- ScanCode powers ORT
- But there is more in AboutCode!
- AboutCode stack roadmap
- Questions?

About me

- On a mission to enable easier and safer to reuse FOSS code with best-in-class open source Software Composition Analysis (SCA) tools, data, and standards for open source discovery, license & security compliance
- Lead maintainer of AboutCode projects (ScanCode, DejaCode, VulnerableCode and others)
- CTO and co-founder of nexB, Inc.
  - pombredanne@nexb.com
  - GitHub: https://github.com/pombredanne
  - LinkedIn: https://www.linkedin.com/in/philippeombredanne
  - Often assisted by Chihuahua Technical Advisor



About AboutCode

AboutCode's FOSS-first mission: FOSS for FOSS

- **Open source tools and open knowledge base (AboutCode stack)**
- Simple and practical standards (Package-URL / PURL: https://github.com/package-url)
- Applications for Legal & Business users (DejaCode) with APIs for everything
- Co-founders of SPDX: https://spdx.org
- Contributors to CycloneDX: https://cyclonedx.org
- Co-founders of ClearlyDefined: https://clearlydefined.io
- Anchors for a community of SCA tools user and developers
- Supported by contributors, nexB and others generous sponsors and supporters!
- nexB provides services and support for SCA to sustain AboutCode FOSS development

Software Composition Analysis (SCA)?

AboutCode

- **Identification** - Identify distinct “units” of third-party software used in a product or project and their provenance
- **Licensing** – Determine the licensing for each software unit
- **Security** – Identify known security vulnerabilities for each software unit
- **Quality** – Evaluate the quality based on development data, such as number of bugs, fixes, etc. - this is the domain of the CHAOSS project
  - Read "SCA the FOSS Way": https://www.nexb.com/software-composition-analysis/
- Everyone is now **their own integrator** of thousand FOSS components
- **A core competency** for any software development organization
- Software Composition Analysis **demands accurate data** for automation
  - Embed in the software development workflow from design through release - as it is in manufacturing

ScanCode powers  the ORT ecosystem

ScanCode is the industry leading code scanner for:

- license detection
- copyright detection

... and is the work horse inside ORT

But also in **ORT**:

- VulnerableCode for **vulnerability** data
- ScanCode LicenseDB for reference **license** data
- **nuget**-inspector for .NET dependency resolution
- **python**-inspector for Python dependency resolution
- Package-URL/**PURL**: Spec and tools for identifying packages (in process to be an ECMA standard)

SCA Tools

But there is more in AboutCode! [1]

- **MatchCode** - code matching for packages and files
  - A different approach for better code matching
  - And soon with **snippets** too
- **ScanCode.io** - scripted, customizable pipelines
  - **Binary analysis** of deployed code (Java, JS), next ELF, see roadmap.
  - **Docker and VM image** analysis
  - Combine scanning and matching analysis in pipeline scripts
- Also in **ScanCode**
  - Comprehensive **package and dependency** manifest scanning
  - License **summarization**, TODO reviews, **license clarity** scoring
- **Standards -** PURL and new **VERS/Univers**
  - Compare package versions ranges - Adopted in OSV, CycloneDX, CSAF

SCA Tools

Management Apps

Open Knowledge Base

But there is more in AboutCode! [2]

- **DejaCode** - new Supply Chain Control Center app
  - System of records for all your products for origin, license and security compliance
  - Ingest, aggregate, manage and export SBOMs across all software
  - Web-based Enterprise management app with SSO, Policies, Reporting, Workflows, access controls, REST API, webhooks
- **Open knowledge base** - licenses, packages & vulnerabilities curated reference
  - **LicenseDB** - open source and other licenses. 2000+ licenses, 35K rules
  - **PurlDB** - reference package data and scans, and code matching - 21M+ packages
  - **VulnerabilityDB** - aggregated and correlated vulnerability data - 760K+ packages and 240K+ vulnerabilities

SCA Tools

Management Apps

Open Knowledge Base

AboutCode

SCA Tools

ScanCode

- Scan
- Match
- Analysis pipelines
- Binary analysis
- Dependency analysis

Management Apps

DejaCode

- Policies
- Curations
- Software inventory
- Workflows
- SBOMs
- Custom reports

Open Knowledge Base

- Licenses
- Packages
- Vulnerabilities

Roadmap for ScanCode Toolkit

- More and better license detection rules
- **Single exe** standalone apps for easier ScanCode deployment in CI/CD
- New lightweight **package-only scanner**
- New summarization and **license clarity scoring** and summarization to reduce noise
- Improve copyright and license detection **speed**
- New models for Packages, Dependencies and Requirements
  - Now with the detection of which files belong to a package
- Move inconclusive, unknown license detection to clues

SCA Tools

Roadmap for ScanCode.io

- Integrate ScanCode.io with **CI/CD** and other tools
  - Create CI/CD pre-configured integrations with main CI (GitHub, GitLab, Jenkins)
- Extend **binary analysis** and tracing workflows
  - Support **ELF/Native, Go**, Python, Ruby, Android in addition to Java and JS
  - Find the exact subset of the code that is deployed and used in production
- **Automate analysis review** in ScanCode.io
  - End to end automated pipelines for embedded devices, Android, C/C++
  - Multi-stack binary deployment analysis for Java, JS, C/C++
  - Report TODO items to review only "by exception"
- Compare scans to focus review work on changes only
- New code inspectors for dependency resolution

SCA Tools

Roadmap for MatchCode code matching

- New Web-based code matching server
- Includes mining to create custom knowledge base
- Match code **snippets** approximately
- Match source and binary symbols to sources and binaries (purl2sym)
- New scripted and customizable matching pipelines
- Accurately match to the correct package version of many possible versions
- Smarter matching in multiple scripted pipeline with ranking to pick best matches

SCA Tools

Open Knowledge Base

Roadmap for Management Apps

- Add support for CycloneDX 1.6 (1.5 is currently available)
- Add support for SPDX 3.0 (2.3 is currently available)
- Create new review automation apps for license and code matches
- **Goal: Zero curation!!!** reduce and automate review work
- New app for advanced **Vulnerability management** and support for CRA (Cyber Resiliency Act) compliance
- Automated triage of vulnerabilities and workflow triggers in your products
- VEX creation, VEX import and export (Vulnerability Exploitability Exchange) with **CSAF** and CycloneDX

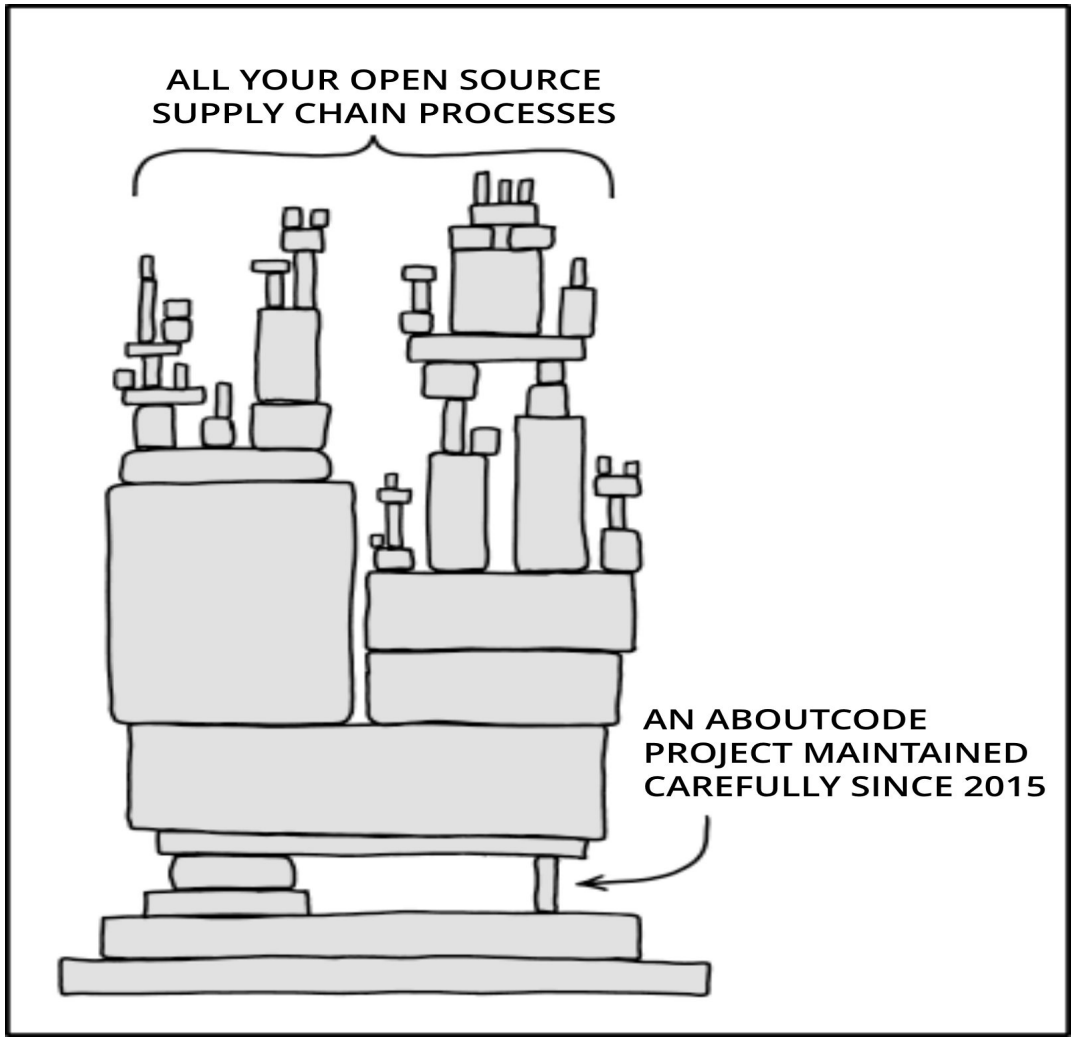
Management Apps

ScanCode and AboutCode need you!

AboutCode

Report to us all your license & copyright bugs!

"Lord Kitchener Wants You" by Alfred Leete is in the public domain / Cropped from original.



"Dependency" by xkcd, used under CC BY-NC 2.5 / Modified text from original

AboutCode

Beyond: Knowledge base

Knowledge base: Packages

- **Packages**: 21M+ packages and, files and their fingerprints
  - PURL-based
  - Public PurlDB is at: https://public.purldb.io/api/packages/
  - All major ecosystems and distributions - sources AND binaries
  - Built-in mining of all package ecosystems, not half-baked
  - Collect, scan, and index package sources, binaries and VCS repos
  - Index with code fingerprints used for code matching
- **Other Package databases**:
  - Software Heritage, ClearlyDefined, deps.dev (Google)
  - Either centralized, proprietary data or code or too big to share on prem
  - No on-premises option for private operations

Open Knowledge Base

Roadmap for Packages [1]

- Confirm the true origin of code to avoid ambiguous matches
- Supply chain package verification
  - Map deployed binary packages to their corresponding source code
  - Find suspicious code drift between package versions
- Mine extensive list of "off registry" packages
  - Common native C/C++ code and libraries for embedded
  - Glibc, Busybox, zlib, etc. not published on ecosystem package registries
- Collect code symbols from source and binaries (for matching)
- Integrate other curation data sources
  - OSSelot
  - ClearlyDefined

Open Knowledge Base

Roadmap for Packages [2]

- On-demand code mining to build your knowledge base
- Federated, decentralized shared knowledge base data with Git and ActivityPub
  - Share scans, vulnerabilities, origin facts, and curations
  - Scan once, analyze once, and collaborate on reviews to clear out the junk!

Open Knowledge Base

Knowledge base: Licenses

- **Licenses**: 2,000+ licenses and 35,000 rules
  - ScanCode LicenseDB holds the core license data
  - ScanCode Toolkit has the license detection rules
  - DejaCode is synchronized with LicenseDB and adds License Conditions
  - All licenses have SPDX Identifiers with “Licenseref-scancode” namespace for the many licenses not included in the SPDX License List (currently only ~700 SPDX licenses)
- No known open alternative with comparable depth and breadth

Open Knowledge Base

Roadmap for Licenses

- Extend License data with compatibility matrix [OSADL, Flict]
- Add new license aliases dataset [@hesa]
- Add more extensive tagging and categorization
- Extend License data with improved exception details
  - To disambiguate license detections of L/GPL with/without exceptions
- Extend License data with improved "or later" details
  - To disambiguate detection of "or later" notices with their primary texts
- Add "key phrases" to all license detection rules
- Add variable text segments to license rules
- Add Fedora alternative SPDX identifiers
- Work with CycloneDX to become their license reference

Open Knowledge Base

Knowledge base: Vulnerabilities

- **Vulnerabilities**: 760K+ packages and 240K+ vulnerabilities
  - PURL-based
  - Public VulnerableCode DB is at: https://public.vulnerablecode.io/
  - All major ecosystems and vulnerability DBs aggregated and correlated
  - Discover relations (and inconsistencies) in data from mining the graph
- **Other vulnerability databases**:
  - OSV (reuses some AboutCode code), GitHub, GitLab, NVD
  - Often contain conflicting data for vulnerable ranges, fixed versions or affected packages

Open Knowledge Base

Roadmap for Vulnerabilities

- Extend non-vulnerable dependency resolution
  - Beyond Python - add Java and JS
- Extend vulnerability data with new upstream data sources
- Add fix commit details, support for vulnerability reachability
- Mine the graph to surface related package fixes
- Mine git logs, issues and forums to enrich vulnerability data
- Surface inconsistencies and conflicts between different advisory data sources (VulnTotal throughout)
- Add source/binary discrepancy data (from back2source)
- Federated, decentralized shared knowledge base data with Git and ActivityPub

Open Knowledge Base