FOSS Software Composition Analysis

Philippe Ombredanne

- ScanCode lead maintainer
- Co-founder of SPDX, ClearlyDefined
- long time GSoC mentor
- Co-founder and CTO of nexB Inc.
- VI/M (still) and Eclipse
- Weird facts and claims to fame
  - Signed off some of the largest deletion of lines of code in the Linux kernel (but these were only comments)
  - Used to have 60K GH forks (now only 20K)
- pombredanne@gmail.com irc:pombreda

The software composition challenges

ScanCode

What's in your code?!

- Ever **more software packages** are reused
  - *10x to 100x more than a few years ago*
- Software origin discovery is still **unsolved**
  - FOSS is so easy to provision and install
- **Clarity in licensing** is far from there
- No single scanning technique and tool is good enough, alone
- Naming and exchanging data about software is hard
- Open and accurate metadata are direly missing

The Vision

ScanCode

On a mission to make it easier and safer to reuse FOSS

- 1. Create tools & libraries for **primary evidence collection**
  - e.g. license, copyright, package manifests, build, etc.
  - Reusable to integrate by other FOSS projects
  - Best-in-class, no compromise detection accuracy
- 2. **Automate composition analysis** in scripted pipelines integrating best-of-breed FOSS tools
- 3. Create **open reference data** sets and models to automate composition analysis

The ScanCode approach

ScanCode

- **Static analysis** as primary technique
- Everything **data driven**
- Use open **metadata database**(s)
- Vet **ALL** the files
- **Complex** composition analysis scripted and customizable
- **Collaboration** for integration in other FOSS projects

Who's using ScanCode

- Used at FOSS orgs and projects
  - BANG, CHAOSS, ClearlyDefined, Eclipse, FSFE, Linux kernel, Object Web, OpenEmbedded.org, Openshift analytics, ORT, Tern and others.
- Used at major companies
  - Accenture, Amazon, BMW, Bosch, Comcast, Facebook, Google, Here.com, Philips, Red Hat, Siemens, Zeiss and others.

The team and the community

ScanCode

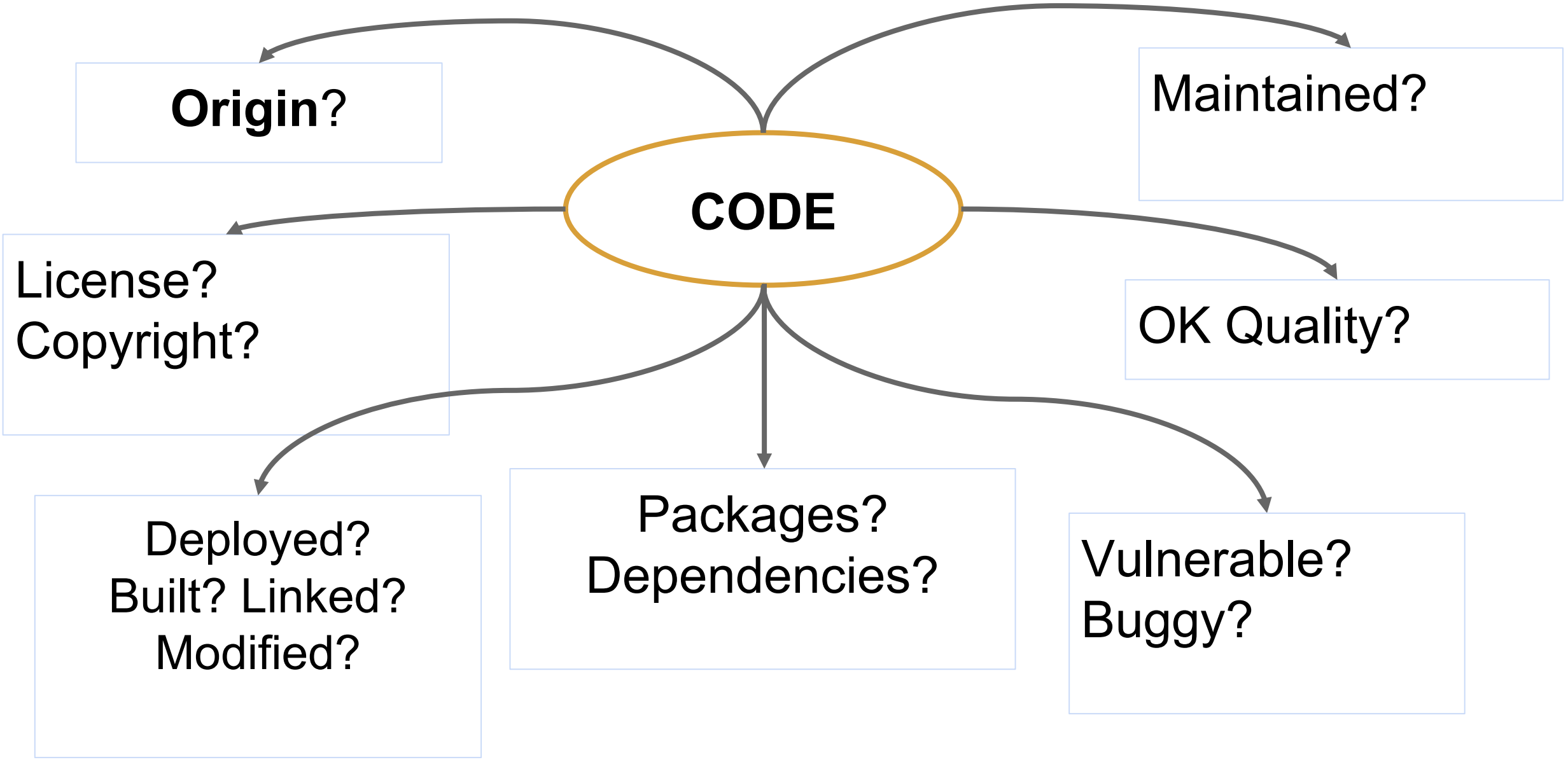
- Over 1,100+ stars @ GitHub
- Over 300 forks
- Over 100 contributors, 700+ chat participants
- Multiple times Google Summer of Code mentoring org
- AboutCode team members are thought leaders
  - Co-founders of SPDX - https://spdx.org
  - Creators of Package URLs - https://github.com/package-url
  - Co-founders of ClearlyDefined - https://clearlydefined.io

Alternatives to ScanCode and AboutCode

ScanCode

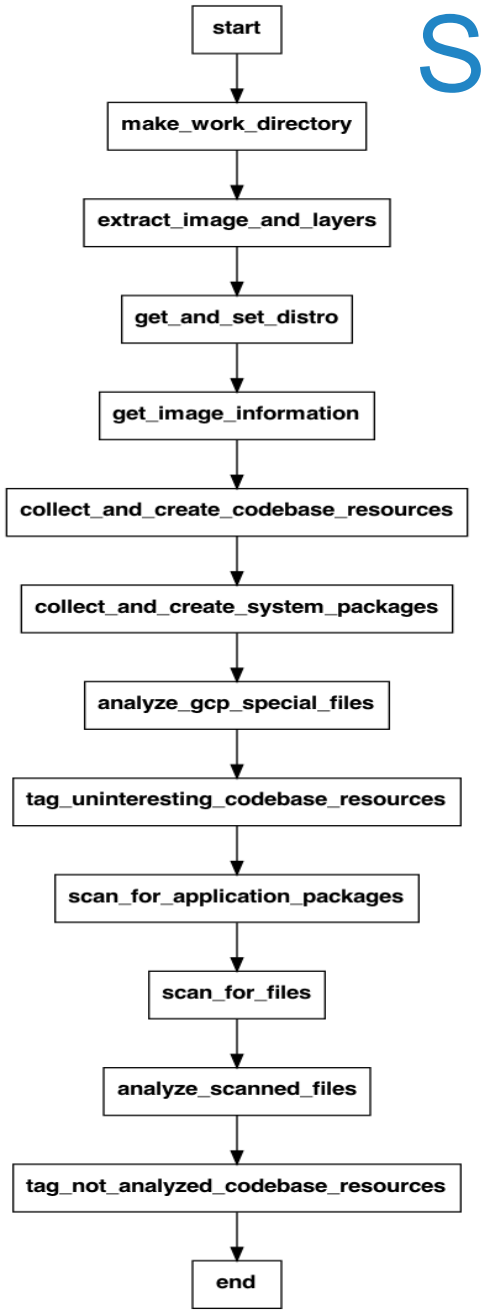
- Commercial tools are focused primarily on Security
  - Weak support for licensing
  - Mostly surface and weak detection of origin
  - Proprietary code, but above all proprietary data
- Open source tools
  - They are presented here today!
  - All are excellent and the more you them the better

Collect many things about the code



ScanPipe Docker SCA pipeline example

ScanCode



- Upload and extract image, find distro OS
- For each image layer: scan **system packages**
  - Find their file and check if modified
- For remaining files: scan **application packages**
  - All ScanCode-supported package types (npm, maven, composer, etc.)
- For remaining files: scan files
  - All files, including binaries
- For remaining files: analyze and tag
  - Dispose of temp and transient or log files and more
- Assemble results from DB and return JSON, XLSX and present web UI

Flagship Projects

- **ScanCode TK**
  - License, Copyrights, Package manifests
  - Each feature is being moved to its own repo/library
  - Used in ORT, Tern, Quartermaster, CHAOS
- **ScanCode.io: composition analysis automation, bespoke SCA**
  - Rest API, database-backed server
  - Emerging Web UI for SCA review
  - Data science-inspired scripted **pipelines** for composition analysis
  - Starting with Docker and VM/rootfs images using static analysis

More projects and libraries

- ScanCode Results Analyzer - Use AI/ML to curate license scans, automagically
- AttributeCode TK - Auto generate attribution notices
- DeltaCode - compare two scans
- Container-Inspector - Static Docker images analysis - low level library
- Debian-Inspector - Debian package manifests parsing
- ScanCode Workbench - Desktop app for Scan review
- license expression - parse, combine, simplify
- Package URL - the new standard Package id used at OWASP, Sonatype, etc.
- TraceCode TK - trace your build to find deployed code
- ExtractCode - uncompress and unarchive all the things
- CommonCode - shared common utilities
- TypeCode - find the type and classify the content of all the files

Reference data and specs

- VulnerableCode - The free correlated DB of ALL the FOSS **vulnerabilities** (with support from the EU and NLnet.nl)
- LicenseDB - **All the licenses** - FOSS or proprietary - https://scancode-licensedb.aboutcode.org/
- ClearCode - Extract all the data and ScanCode scans from ClearlyDefined
  - Deployed also at Software Heritage
- Upcoming:
  - PackageDB - **All the packages** and dependencies metadata

- Data models and specs
  - AboutCode - Data models (used in Libraries.io and ORT)
  - **Package URL** (purl) - Mostly universal Package identifier - used in OWASP, Sonatype

Plans and next...

ScanCode

- **MOAR DATA** ABOUT CODE!
  - **PackageDB** as open reference FOSS data
  - Build a graph of all packages, licenses, vulnerabilities
  - Scan and review accuracy of scans of **all the FOSS code**
  - **More licenses** in the LicenseDB (1.6K today) and more notices samples (20K today)
- **MOAR CODE** ABOUT CODE!
  - **Matching** against index of software package and files
  - **AI/ML-assisted data curation** (and clearing)
  - More **automation pipelines** of the core SCA
- MOAR Documentation too!

Credits

ScanCode

Special thanks to all the people who made and released these excellent free resources:

- Presentation template by SlidesCarnival
- Photographs by Unsplash
- All the open source software authors that made ScanCode and AboutCode possible