



# SCANOSS Journey

From Open Source Software...  
towards Open Data

Julian Coccia  
SCANOSS CTO  
FOSSNorth  
April 2024

<http://scanoss.com>  
<http://www.github.com/scanoss>

# ToC

- About SCANOSS
- Our journey: from Open Source software...
- Our journey: ... towards open data
  - OSSKB
  - Purl2cpe
  - Crypto\_algoritms\_open\_dataset
- Next steps
- Summary



# About SCANOSS





# About **SCANOSS**

- Software Composition Analysis disruptor
- Registered in Madrid, Spain, EU
- 30+ team members (2024Q1)
- Key executives
  - Alan Facey (Co-founder and CEO)
    - Investor, board member, chairman/advisor of several companies
    - Former CRO at Black Duck Software
  - Julian Coccia (CTO)
    - FOSSID co-founder and former CTO
    - Former FOSS Expert at Ericsson/ST-Ericsson



# Software composition market

- Started as a proprietary monopoly
- OSS grew but that monopoly remained
- SCANOSS challenges the status quo with an OSS business model
- Is it feasible to disrupt this market with such business models? **YES**
- Is it feasible to do it with business models based on open data? **YES**



# About US

- A **data** company
- **IP:** Knowledge Base and Mining Network
- Knowledge Base of **published** OSS
  - **No** proprietary software
  - **No** unpublished OSS





# Building and maintaining a knowledge base

- The Knowledge Base is BIG
  - >2Pb downloaded
  - 210M URLs
  - 100B files
  - 3T lines of code
- Creating and maintaining a KB is expensive
  - Dedicated team of data scientists/engineer
  - Dedicated team of curators
- Providing a reliable data access at scale is expensive
  - Dedicated operations team
  - High hosting costs



# Free SBOM with Undeclared Software detection

Detecting declared software	<ul style="list-style-type: none"><li>• Reading software metadata</li><li>• Copyright statements</li><li>• License files and headers</li><li>• Declared dependency files</li></ul>
Detecting undeclared software	<ul style="list-style-type: none"><li>• Files without license headers</li><li>• Built-in dependencies</li><li>• Stripped headers</li><li>• Plagiarized code</li><li>• AI-generated code containing OSS</li></ul>





# Key paid features

- Open Source intelligence
  - Export control, quality, provenance, sustainability, etc
- Matching/diffs between upstream and your open source code
- High precision snippet matching (HPSM)
  - Increased resolution, reduced false positives
- Guaranteed availability and throughput
  - SaaS and On-Prem



# Our Journey: from Open Source Software...



# About US

- SCANOSS makes Open Source Software
- Our Knowledge Base is made using OSS only (SCANOSS, Scancode, Semgrep, etc.)
- Some of our contributions:
  - **Minr**: Make your own Open Source Knowledgebase
  - **Engine**: Scan against your Knowledgebase
  - **API**: OpenAPI / Protobuf
  - **SDKs**: Python, Java, Javascript
  - **CLIs**: Extract fingerprints and scan against Knowledgebase
  - **UI**: SBOM Workbench (multiplatform app)
  - **Plugins** and extensions to integrate various tools and the DB



OUR JOURNEY... FROM OPEN SOURCE SOFTWARE

# Our Memberships

- **OpenChain** Partner
  - OpenChain Tooling WG
  - OpenChain Export Control WG
- **Eclipse Foundation** Member
  - Eclipse SDV Member
- **Software Heritage** Sponsor
- **OSPO Alliance** Member
- Others will come in this 2024



OUR JOURNEY... FROM OPEN SOURCE SOFTWARE

# Growing Ecosystem

## Open Source SCA



+ others

## Auditing firms



## Courts

## Commercial SCA vendors



## Universities




# Our Journey: ...towards Open Data





# Core Contributions

1. Input and algorithms
2. Free access to Open Source Software Knowledge Base (OSSKB)   
<https://osskb.org>
3. Outputs: published open data sets
  1. 2022 Launched PURL2CPE  
<https://github.com/scanoss/purl2cpe>
  2. 2024 Launched crypto\_algorithms\_open\_dataset  
[https://github.com/scanoss/crypto\\_algorithms\\_open\\_dataset](https://github.com/scanoss/crypto_algorithms_open_dataset)
  - More this year!





# 1. Input and algorithms

- The data SCANOSS collects, store, process and curate to create its Knowledge Base is published open source and public domain software only.
  - Our knowledge base does not include data from our customers
- The algorithms used by SCANOSS are open source
  - Winnowing, snippet fingerprinting
  - Minr, mining and cryptography detection
- Our own implementations of algorithms are published under permissive open source licenses.





## 2. Access to SCANOSS KB: OSSKB

- In 2021, SCANOSS has provided to the Software Transparency Foundation (STF) a perpetual license to host and provide free (gratis) access to OSS KB
- OSSKB is a subset of SCANOSS KB.
- It targets open source developers, tools and researchers so they can:
  - Detect declared and undeclared OSS in their SW products
  - Check their software against plagiarism introduced by copy/paste or by AI-assistants
  - Integrate in their pipelines/workflows basic software composition analysis activities



## 2. Access to SCANOSS KB: OSSKB

- Use OSSKB <https://osskb.org>
  - STF prevents abusive usage of this OSSKB guaranteeing a reasonable QoS for all.
- Contribute to OSS KB. We encourage organizations to join the Software Transparency Foundation to help us to:
  - Promote access to existing and future free (gratis) commercial and non-commercial data sets based on public, open data, solving inbound SBOMs
  - Finance to make those data sets consumable at scale by creators, researchers, open source tools, etc.
  - Mirror OSS KB to guarantee QoS to your own users: NGOs, education institutions, non-profits...



## 3.1 purl2cpe

- Dataset with relations between PURLs and CPEs
- Published as Open Data

<https://github.com/scanoss/purl2cpe>

- Output of a team of curators. Daily updates
- Easy to consume, easy to contribute
- Useful to automatically check for vulnerabilities in your components



## 3.1 purl2cpe

- How to use it
  - A SQLite database is automatically created from the data/ directory
  - Easy to download periodically and query your PURLs for vulnerabilities
- How to contribute
  - Create your issues or PRs in the Github repo

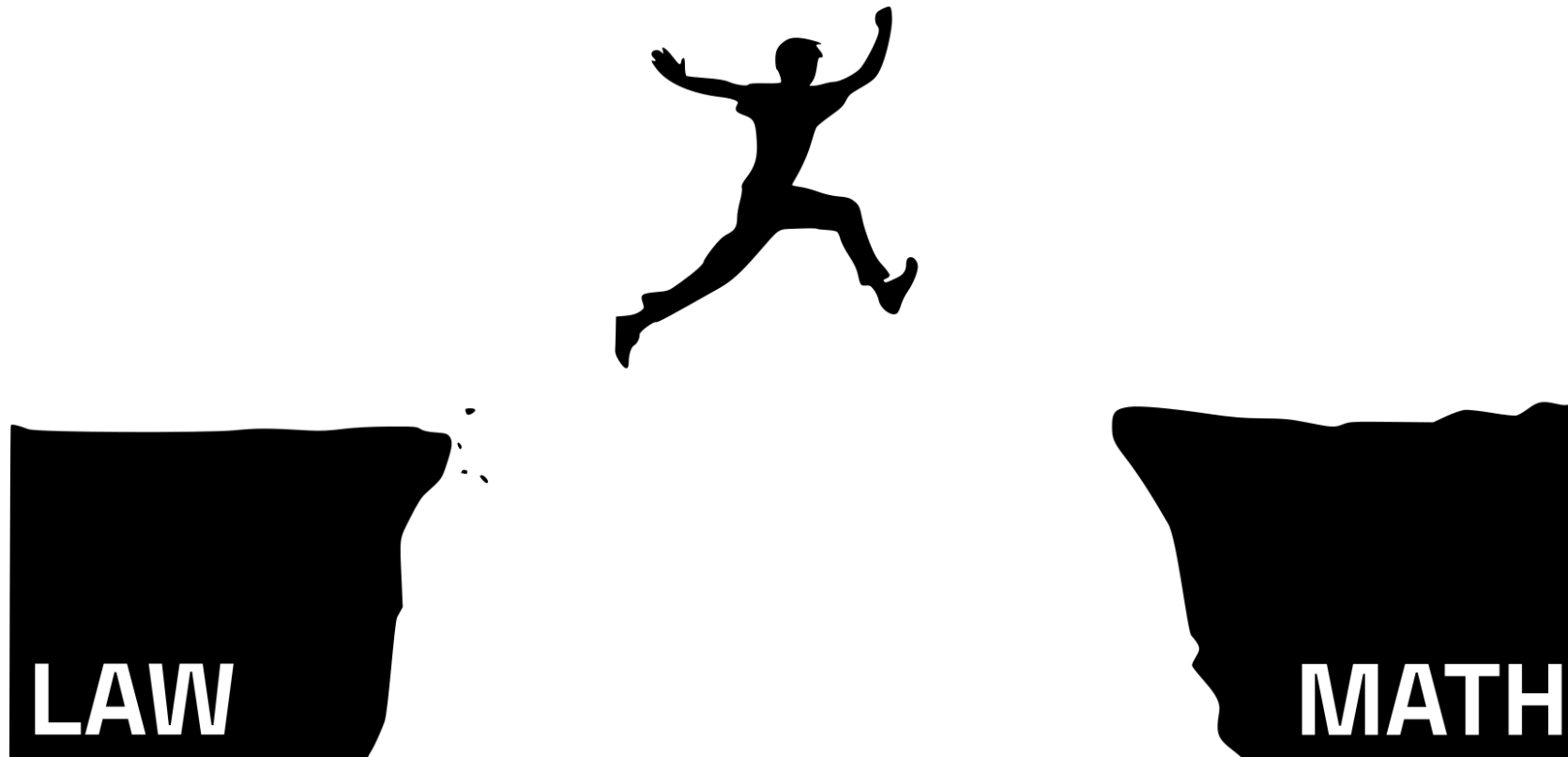
<https://github.com/scanoss/purl2cpe>

- Follow us to stay up to date!



OUR JOURNEY... TOWARDS OPEN DATA

## 3.2 Crypto algorithms & ECCN: bridging the gap together





## 3.2 Crypto algorithms definitions goal: consensus

- Just like we (as communities) did with license compliance...
  - Reducing overall **costs and risks**
  - Improving **transparency and efficiency**
- How to **declare** these open source algorithms
  - Initial step towards **standardization**
  - Algorithm **classification and attributes**
- Bringing declarations **into SBOMs**
  - Simpler distribution, consumption and management.



## 3.2 crypto\_algorithms\_open\_dataset: definition

The screenshot shows a GitHub repository page for 'Toscalix crypto algorithms (#4)'. The repository was created 3 weeks ago and has 8 commits. The file list includes:

File Name	Commit Message	Commit Date
.idea	Initial commit	last month
definitions_crypto_algorithms	Toscalix crypto algorithms (#4)	3 weeks ago
docs_crypto_algorithms	Toscalix crypto algorithms (#4)	3 weeks ago
utilities	Toscalix crypto algorithms (#4)	3 weeks ago
.gitignore	Initial commit	last month
CODE_OF_CONDUCT.md	Initial commit	last month
CONTRIBUTING.md	Initial commit	last month
LICENSE	Initial commit	last month
README.md	Toscalix crypto algorithms (#4)	3 weeks ago

Navigation links: README, Code of conduct, CC0-1.0 license.

### Cryptographic Algorithms Open Dataset

This data set, which includes a list of cryptography algorithms with an open source implementation, was originally the output of SCANOSS mining efforts across its entire data base, which includes all relevant open source software published. Today, the intention is to turn this repository into a collaborative project to enrich and maintain this data set, not just for export control, the original target activity, but for other purposes as well, like quantum safe or compliance with a variety of regulations.





## 3.2 crypto\_algorithms\_open\_dataset: definition

This is the index of the cryptography algorithms definitions including the algorithmId and the link to the definition of each algorithm.

- Date: 2024-03-14
- listVersion:

The table sorts by the algorithmId column.

#	Cryptography Algorithm Name	algorithmId	Link to Definition
1	3des	3des	<a href="#">Link</a>
2	3way	3way	<a href="#">Link</a>
3	ASN1	ASN1	<a href="#">Link</a>
4	CMAC	CMAC	<a href="#">Link</a>
5	X509	X509	<a href="#">Link</a>
6	aes	aes	<a href="#">Link</a>
7	aria	aria	<a href="#">Link</a>
8	bcrypt	bcrypt	<a href="#">Link</a>
9	blakex	blakex	<a href="#">Link</a>
10	blowfish	blowfish	<a href="#">Link</a>
11	blum-goldwasser	blum-goldwasser	<a href="#">Link</a>

Algorithm list, detection definitions, attributes, reference code





## 3.2 crypto\_algorithms\_open\_dataset: use cases

- Export Control (ECCN, Trade Compliance)
- Security compliance
  - Crypto Algorithm Validation Program (CAVP)
    - NIST-recommended
    - FIPS-approved
- Quantum safe area

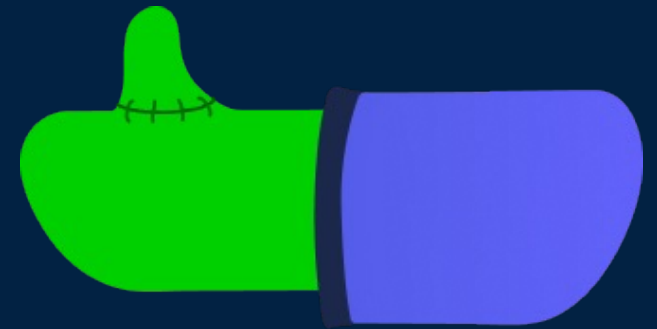


## 3.2 crypto\_algorithms\_open\_dataset: contribute

- Check it out, it is open data.
  - Detect your cryptography now!
- Help your supply chain
  - Use the definitions in SBOMs
  - Add detection capabilities to your tool
- Contribute!
  - New Definitions
  - New Attributes
  - Evolve the specs



# Our Journey: ...towards Open Data





# Open Source software journey: Next steps

From published open source software to open source software projects

- Maintain the software we have already published (we use it!)
- Keep publishing under OSS licenses the software we develop for our commercial offering
- Mature our current contribution policy while keeping it extremely lightweight
- Consolidating and increasing our users base increases the impact of our contributions



# Open Data journey: Next steps

OSSKB:

- Keep it current and healthy
- Keep developing software to access and consume it
- Add new sources

Keep supporting STF to:

- Increase the number of organizations that host a mirror of OSS KB
- Maintain the current levels of QoS and increase them when possible.
- Increase the number of sponsors (especially corporations) so OSS KB supports a larger number of concurrent users and a more intensive usage.



# Open Data journey: Next steps

Already published open data sets

- Enrich them, keep them current and healthy
- Consolidate them as open data collaborative projects or donate them to an open governance organization to boost contributions

and based on our capacity...

- Publish and maintain additional data sets as open data.
- Collaborate in open governance ecosystems providing expertise, publishing and co-maintaining data sets



# Summary



# You can't protect or comply with what you cannot see







# Standardizing how we declare artifacts is like giving glasses to Frankie



## SUMMARY

# From OSS through Open Data to Open Standards



## SUMMARY



# More about SCANOSS

Web: <https://scanoss.com/>

Our software: <https://github.com/scanoss>

OSSKB: <https://osskb.org/>

Our published data sets:

Purl2cpe:

<https://github.com/scanoss/purl2cpe>

Crypto\_algorithms\_open\_dataset:

[https://github.com/scanoss/crypto\\_algorithms\\_open\\_dataset](https://github.com/scanoss/crypto_algorithms_open_dataset)





# SCANOSS Journey

From Open Source Software...  
towards Open Data

Julian Coccia  
SCANOSS CTO  
FOSSNorth  
April 2024

<http://scanoss.com>  
<http://www.github.com/scanoss>