Software Heritage

What a graph can tell you about your software supply-chain ?

Agenda

What is the Software Heritage

Some usages of currently archived data

What we are working on

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

We are building the « Library of Alexandria » of software source code

Mission to collect, preserve, and share all software that is publicly available in source code form

A non profit organization

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Mission: collect, preserve and share all software source code

One infrastructure
open and shared

Started at Inria in 2016

Inria

The largest archive ever

https://www.softwareheritage.org/

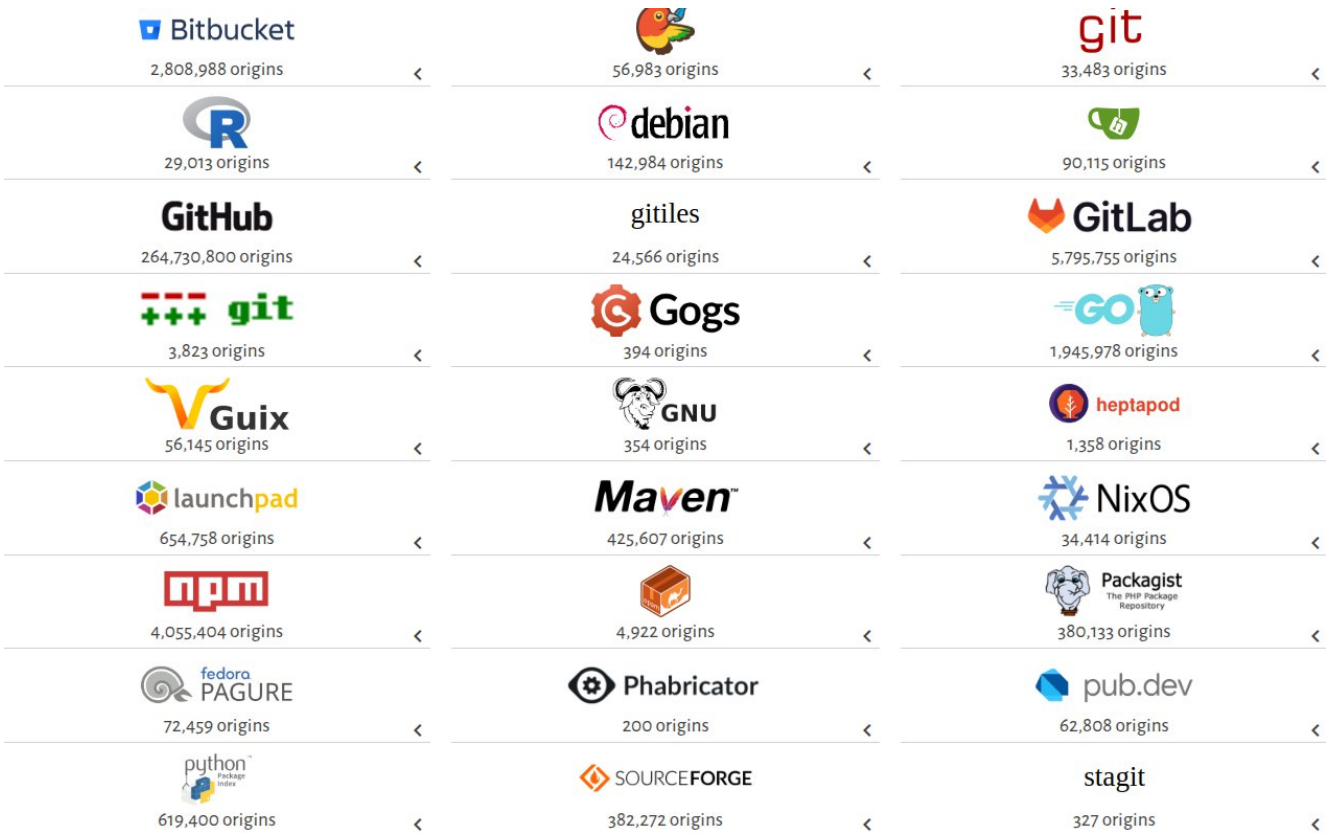
Team and partners

~ 20 peoples + many partners

And several others listed on
https://www.softwareheritage.org/

Data we collect

From a lot of origins:



Some of them do not exist anymore:

Data we collect

- Source code files: 23B deduplicated source file
- History: 5B of code revisions, with date, author, message, related files...

https://github.com/torvalds/linux

31 March 2025, 00:26:56 UTC

Code | Branches (1) | Releases (878) | Visits

Branch: HEAD | 122868a / kernel / cpu.c

Tip revision: **4e82c87058f45e79eeaa4d5bcc3b38dd3dce7209** authored by **Linus Torvalds** on **31 March 2025, 00:03:26 UTC**
Merge tag 'rust-6.15' of git://git.kernel.org/pub/scm/linux/kernel/git/ojeda/linux

cpu.c

```c
/* CPU control.
 * (C) 2001, 2002, 2003, 2004 Rusty Russell
 *
 * This code is licenced under the GPL.
 */
#include <linux/sched/mm.h>
#include <linux/proc_fs.h>
#include <linux/smp.h>
#include <linux/init.h>
#include <linux/notifier.h>
#include <linux/sched/signal.h>
#include <linux/sched/hotplug.h>
#include <linux/sched/isolation.h>
#include <linux/sched/task.h>
#include <linux/sched/smt.h>
#include <linux/unistd.h>
#include <linux/cpu.h>
#include <linux/oom.h>
#include <linux/rcupdate.h>
#include <linux/delay.h>
#include <linux/export.h>
#include <linux/bug.h>
#include <linux/kthread.h>
#include <linux/stop_machine.h>
#include <linux/mutex.h>
#include <linux/gfp.h>
#include <linux/suspend.h>
#include <linux/lockdep.h>
#include <linux/tick.h>
#include <linux/irq.h>
#include <linux/nmi.h>
#include <linux/smpboot.h>
#include <linux/relay.h>
#include <linux/slab.h>
#include <linux/scs.h>
#include <linux/percpu-rwsem.h>
```

https://github.com/torvalds/linux

visit type: git

18 March 2025, 07:35:20 UTC

Code | Branches (1) | Releases (877) | Visits

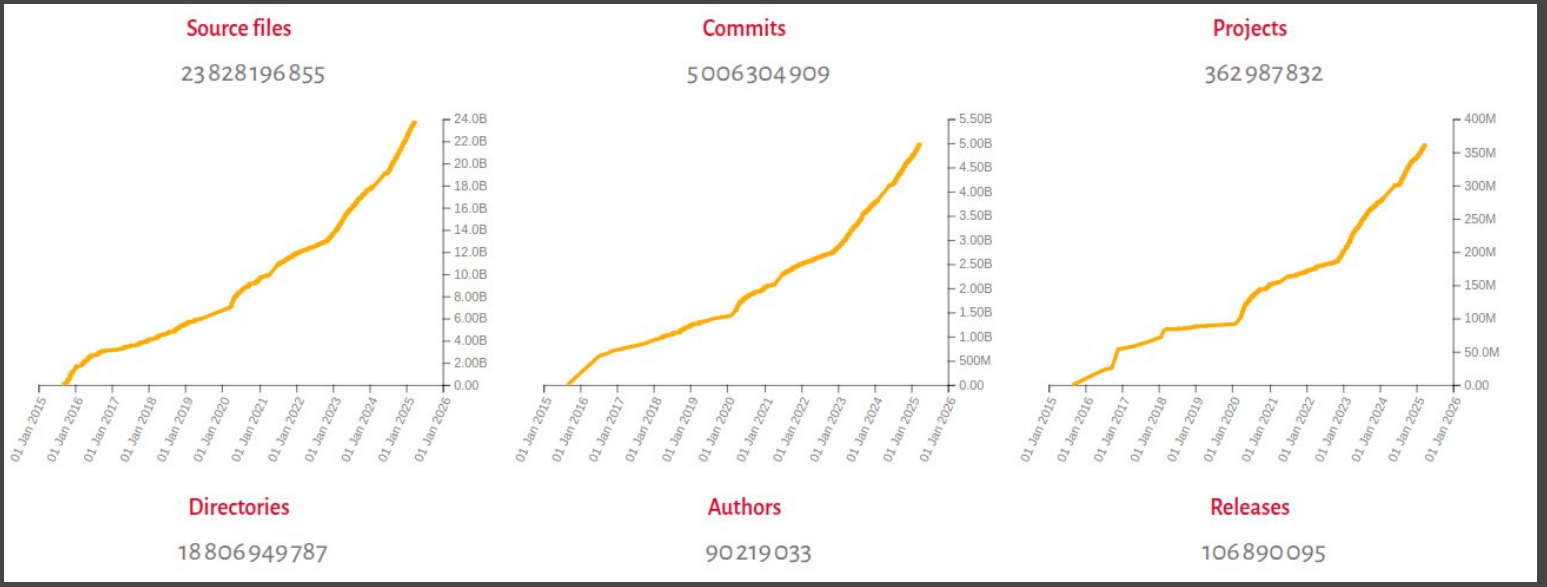
Branch: HEAD | Save again | Extrinsic metadata

sort by: revision date | DFS | DFS post-ordering | BFS

Revision	Author	Date	Message	Commit Date
76b6905	Linus Torvalds	18 March 2025, 05:27:27 UTC	Merge tag 'mm-hotfixes-stable-2025-03-17-20-09' of git://git.kernel.org/pub/scm...	18 March 2025, 05:27:27 UTC
9130945	Eric W. Biederman	17 March 2025, 13:47:30 UTC	MAINTAINERS: Remove myself Unfortunately I no longer have time to meaningf...	18 March 2025, 05:20:48 UTC
fc444ad	Linus Torvalds	17 March 2025, 21:40:40 UTC	Merge tag 'soc-fixes-6.14-2' of git://git.kernel.org/pub/scm/linux/kernel/git/soc/so...	17 March 2025, 21:40:40 UTC
47c7efa	Linus Torvalds	17 March 2025, 21:30:31 UTC	Merge tag 'probes-fixes-v6.14-rc6' of git://git.kernel.org/pub/scm/linux/kernel/git...	17 March 2025, 21:30:31 UTC
800f105	Kirill A. Shutemov	10 March 2025, 08:28:55 UTC	mm/page_alloc: fix memory accept before watermarks gets initialized Watermar...	17 March 2025, 00:40:26 UTC
b9c0e49	Matthew Wilcox (Oracle)	10 March 2025, 14:35:24 UTC	mm: decline to manipulate the refcount on a slab page Slab pages now have a ref...	17 March 2025, 00:40:26 UTC
9f01b49	Shakeel Butt	10 March 2025, 23:09:34 UTC	memcg: drain obj stock on cpu hotplug teardown Currently on cpu hotplug teard...	17 March 2025, 00:40:25 UTC
14efb47	Zi Yan	10 March 2025, 15:57:27 UTC	mm/huge_memory: drop beyond-EOF folios with the right number of refs When ...	17 March 2025, 00:40:25 UTC
67a2f86	Rafael Aquini	18 February 2025, 19:22:51 UTC	selftests/mm: run_vmtests.sh: fix half_ufd_size_MB calculation We noticed that ...	17 March 2025, 00:40:25 UTC
182db97	Raphael S. Carvalho	24 February 2025, 14:37:00 UTC	mm: fix error handling in __filemap_get_folio() with FGP_NOWAIT original repor...	17 March 2025, 00:40:25 UTC
73f839b	Muchun Song	06 March 2025, 02:31:33 UTC	mm: memcontrol: fix swap counter leak from offline cgroup Commit 6769183166...	17 March 2025, 00:40:24 UTC
8c6ff7f	Dev Jain	06 March 2025, 06:30:37 UTC	mm/vma: do not register private-anon mappings with khugepaged during mma...	17 March 2025, 00:40:24 UTC
d7147a3	Zhiyu Zhang	06 March 2025, 13:28:55 UTC	squashfs: fix invalid pointer dereference in squashfs_cache_delete When mounti...	17 March 2025, 00:40:24 UTC
60cf233	Zi Yan	05 March 2025, 20:04:03 UTC	mm/migrate: fix shmem xarray update during migration A shmem folio can be ei...	17 March 2025, 00:40:24 UTC
cb402bb	Jinjiang Tu	04 March 2025, 13:21:06 UTC	mm/hugetlb: fix surplus pages in dissolve_free_huge_page() In dissolve_free_hu...	17 March 2025, 00:40:23 UTC

Data we collect

- Source code files: 23B deduplicated source file
- History: 5B of code revisions, with date, author, message, related files...
- Metadata:
  - From the code itself: codemeta.json
  - From the ecosystem around
    - GitHub stars, declared languages
- Releases, specific versions tagged by developers
- Packages
- ...



How we archive ?

- Bots “crawling” software forges and translating to our data model
- Deposit, from Libraries advocating for Open Science
  - HAL in France
  - Zenodo
  - eLife
  - IPOL Journal
- SWHAP: Software Heritage Acquisition Process for older historical source code

- With some technical gems:
  - Object Storage, to store reliably 23B files representing 1.5PB of data
  - Graph compression algorithm, to be able to access to the whole code history from one server
  - ...

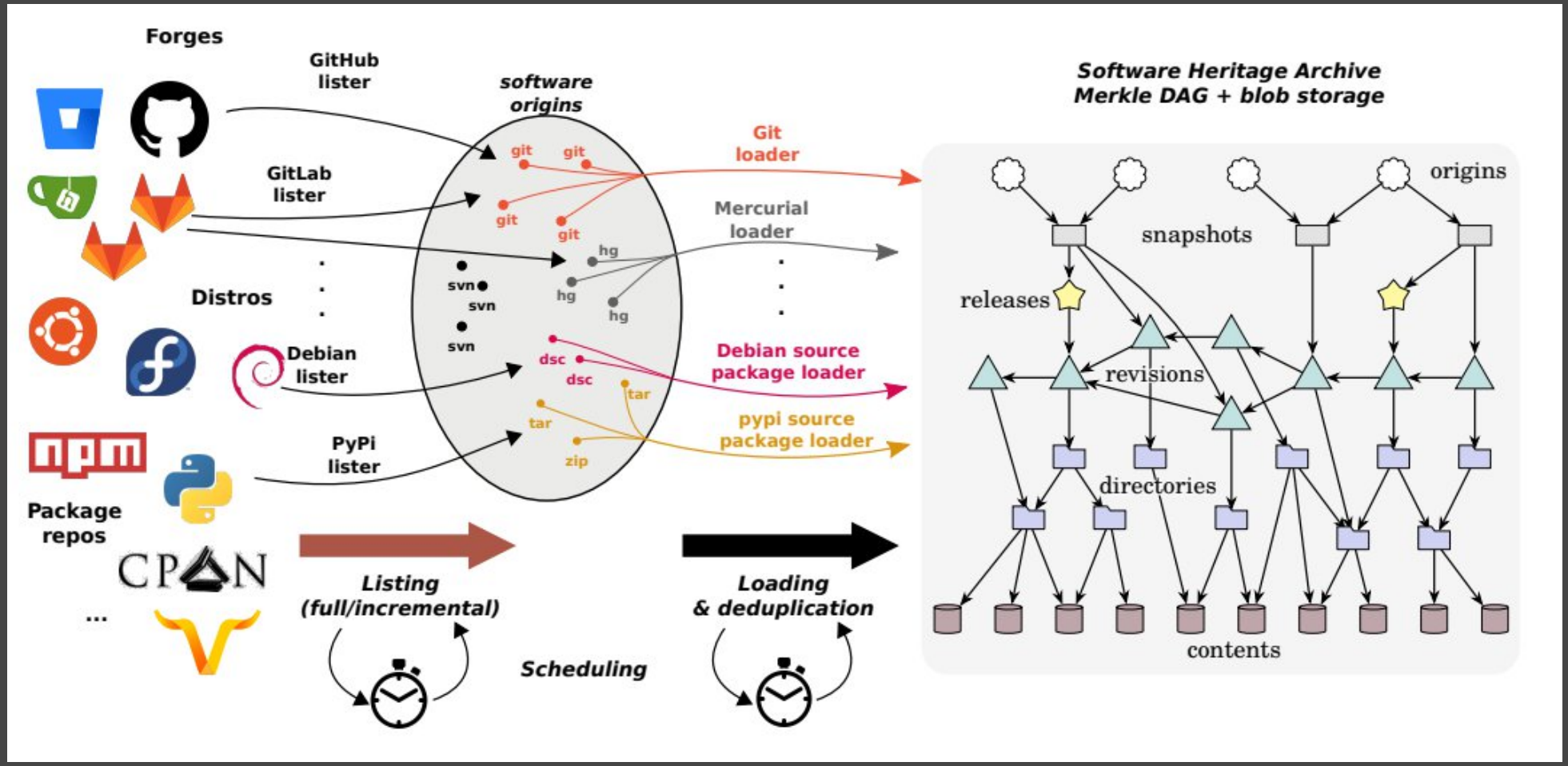
- https://archive.softwareheritage.org/ if you want to browse the archive content
- https://gitlab.softwareheritage.org/ if you want to see archive code

SWHID

Swhid.org

- Software Hash persistent Identifier
- 50+B intrinsic, decentralised, cryptographically strong identifiers, SWHIDs
- Normalization in progress

In a nutshell



- Global development history, permanently archived in a uniform data model
- One infrastructure, shared: more efficient, less waste
- Universal knowledge base for software compliance

Some usages of this archive

Beyond conservation

Science about Software

- A shared science infrastructure
- Datasets
  - Filenames
  - Gender Differences in Public Code Contributions
- Open Science
- Security

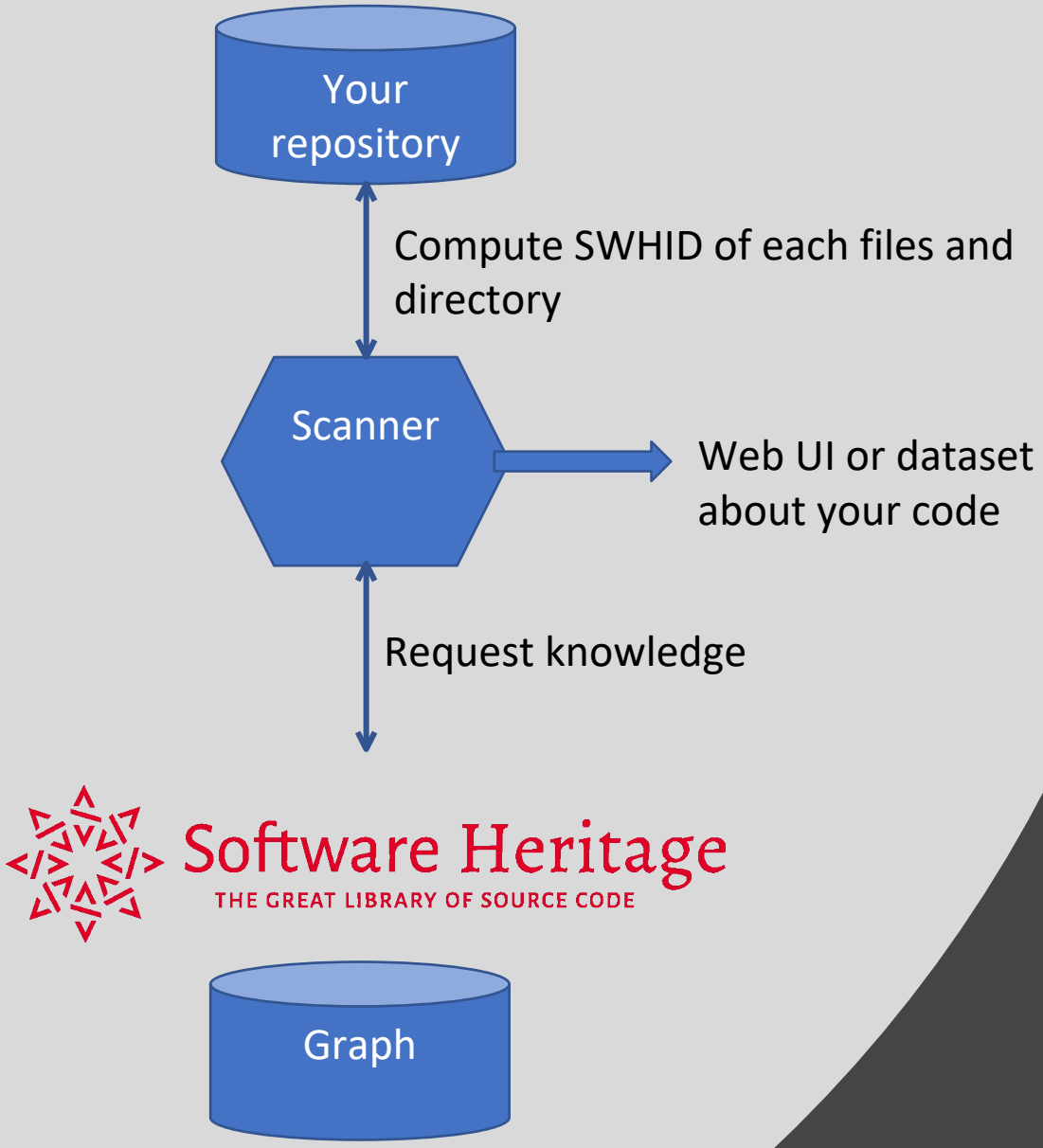
- www.softwareheritage.org/publications/

Complete Corresponding Source compliance

- GPL requires that you provide complete corresponding source

- A potential story:
  - CCS tarballs published at release time, URLs included in user manual
  - Then, a reorganization
    - 404 on CCS Urls -> out of compliance

- A better approach:
  - Prepare CCS tarball
  - Deposit it to Software Heritage
  - Include SWHID in user manuals



Does Software Heritage know my code?

- The scanner
  - open source and open data source code scanner for compliance workflows, backed by the largest public archive of public source code

- Design
  - Software Heritage archive as source of truth about public code
  - Graph and SWHIDs for maximum scanning efficiency
  - File level granularity
  - Output: source tree partition into know (i.e published before) vs unknown

Is my code known by SWH ?

- Ran the scanner on a personal repository
  - open source but altered with a “malicious” file

Dashboard

Results summary

Full known directories	Partially known directories	Files known
87%	12%	96%
7 full known directories out of a total of **8**	**1** partially known directories out of a total of **8**	**24** known files out of a total of **25**

Scanner configuration

Authentication	**Authenticated**
Disable global exclusion patterns	**False**
Disable DVCS exclusion patterns	**False**
Exclusion templates	**[]**
Exclusion patterns	**[]**

Is my code known by SWH ?

- .devcontainer
- .gitignore
- .travis.yml
- .vscode
- CITATION
- LICENSE
- MANIFEST.in
- README.rst
- bin
- community
- docs
- malicious
- requirements.txt
- setup.py
- test_community.py

Results tree

Click on a directory or a file for result details.

setup.py

Definition

Name	setup.py
Type	**content**
Known	✔ **true**
Swhid	swh:1:cnt:87dd4e7f9f8f0076da41987f2c46713b9f512622

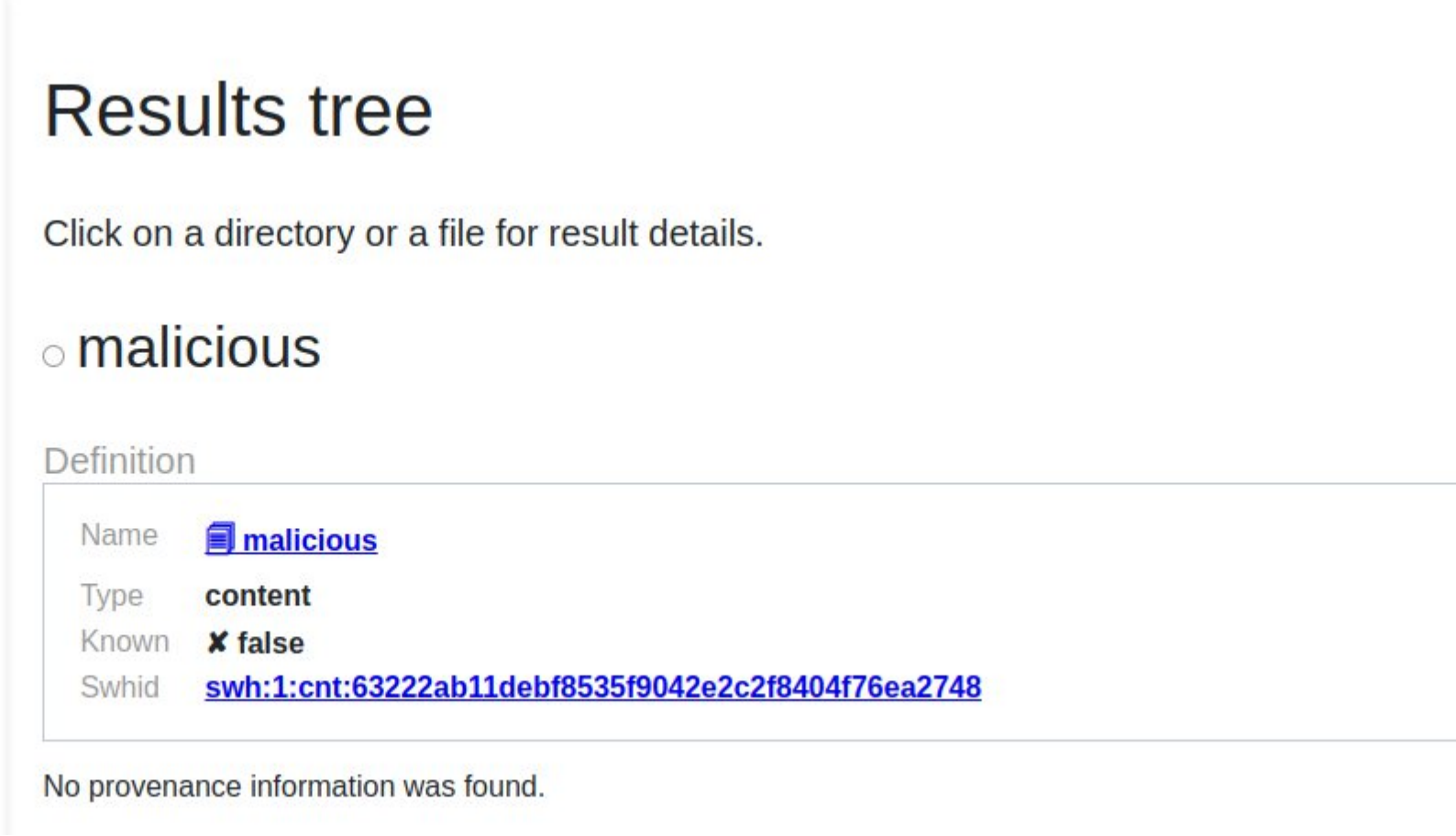
Origin

Url	deb://Debian/packages/python-louvain

Release

Name	**0.16-1**
Swhid	swh:1:rel:b06cf06549120c41be9110bc098f0efeb4b33d2f
Message	**Synthetic release for Debian source package python-louvain version 0.16-1**
Date	**2024-08-06T12:05:58+02:00**
Author	**tchet@debian.org** **Alexandre Detiste** **Alexandre Detiste <tchet@debian.org>**

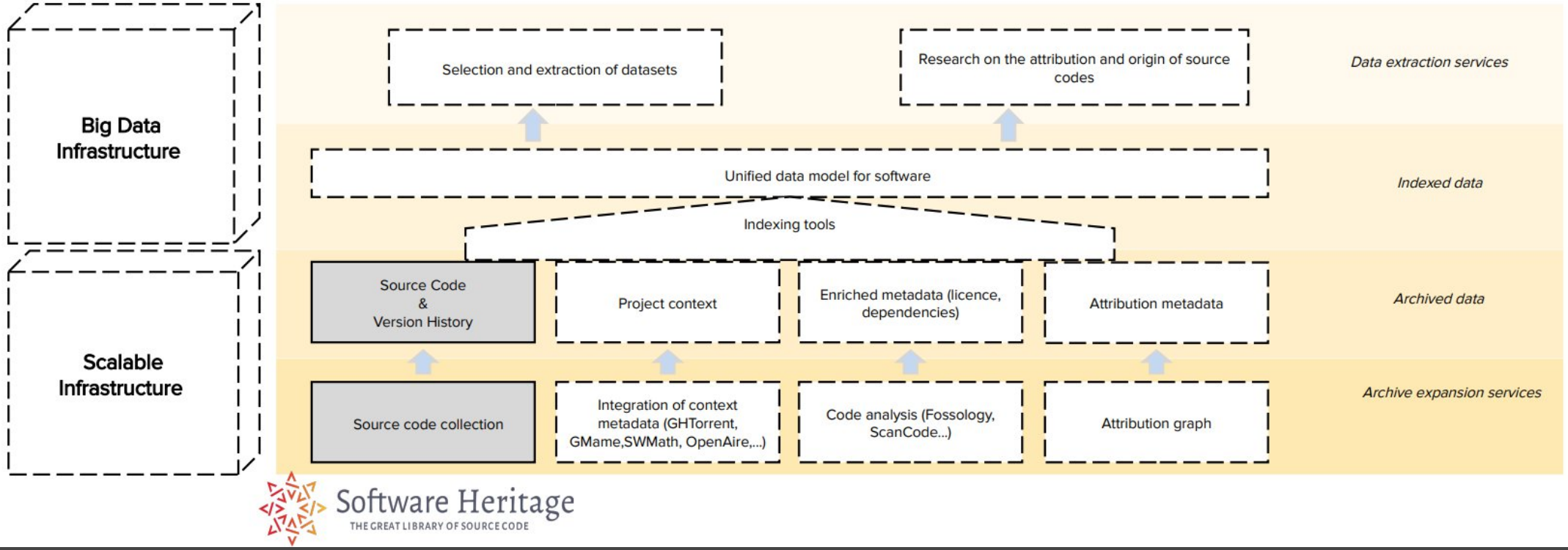
Is my code known by SWH ?



- Do you use open source code ?
- Is one of your open source dependencies altered ?
- Is my private code known on the Internet ?

New work in progress

CodeCommons



- Project with many collaborations to:
  - Archive more data
  - Collect more metadata
    - License, Language, Versions, dependencies Detection
    - Interaction with forges (issues, merge requests, comments...)
  - Detect similar code

CodeCommons

- Project with many collaborations to:
  - Archive more data, faster
  - Collect more metadata
    - License
    - Language Detection
    - Version Detection
    - Dependencies Detection
    - Interaction with forges (issues, merge requests, comments...)
  - Detect similar code

- Build LLMs training datasets
  - A self service shop, where you can ask “most important codes in python or golang, updated in the last two years, with this pool of licenses, no known vulnerabilities and used by academics”
- Massive, transparent and and shared infrastructure with traceability of training data and code attribution

- Extend scanner, license of the files my code depends on ?
- Building SBOM

- https://codecommons.org/

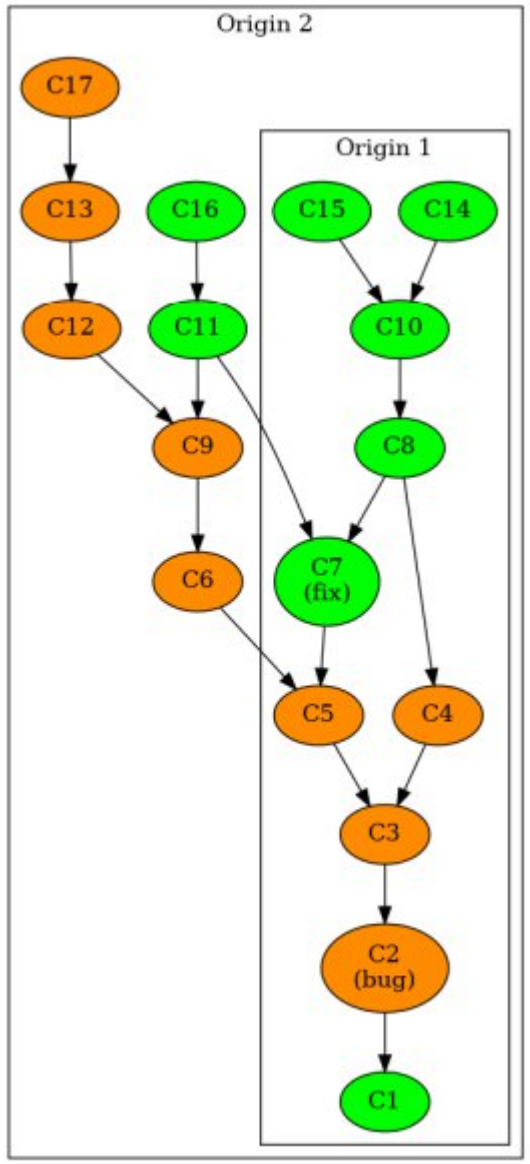
SWH-Sec

- “The Common Vulnerabilities and Exposures (CVE) system provides a reference-method for publicly known information-security vulnerabilities and exposures.”
  - e.g., CVE-2014-0160 (AKA: Heartbleed), CVE-2021-44228 (AKA: Log4Shell)
- Osv.dev: a distributed vulnerability database for Open Source
  - Service (operated by Google) and data format that:
    - Crawls vulnerability information (GitHub, distros, package manager,...)
    - APIs to query the information

- Is my code affected by a known CVE?

```
$ curl -X POST -d \
  '{"commit": "6879efc2c1596d11a6a6ad296f80063b558d5e0f"}' \
  "https://api.osv.dev/v1/query"
{"vulns":[{"id":"OSV-2020-484","summary":"Heap-buffer-overflow in AAT...
```

- But does not know about all code



SWH-SEC and Scanner

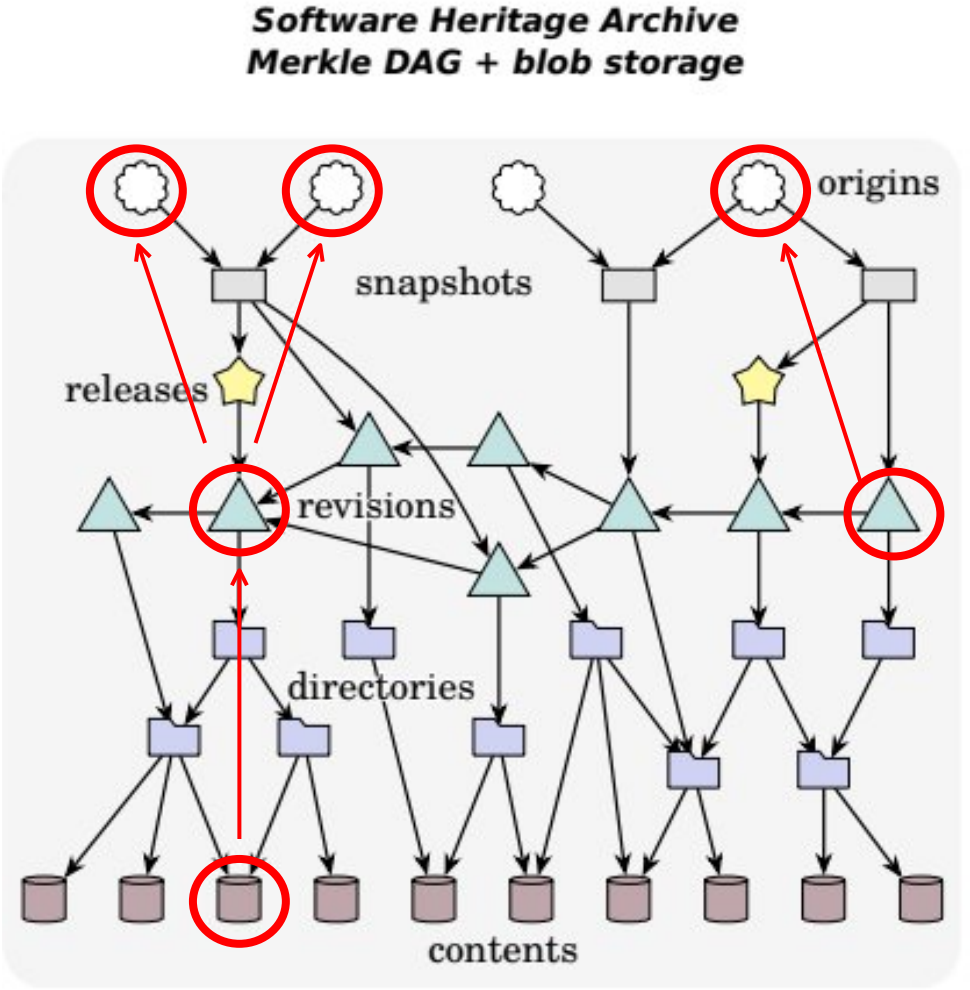
- Project with many collaborations to :
  - Collect CVE
  - Link CVE to files in the history graph
  - Extend the graph with vulnerabilities
- Research work in progress
- After that, maybe
  - Evolution of the scanner to add some features
    - Run it on your infrastructure
    - Run in Continuous Integration
    - Detect some forbidden files (log4j)
    - Better provenance
    - Licenses
    - ?
  - Evolution of the scanner to integrate vulnerabilities
- https://swhsec.github.io/

Cyber Resilience Act obligations

- CRA
  - Security obligations for products with digital content put on the market in Europe

- What Software Heritage brings to the table for Open Source
  - Long term availability (archive)
  - Integrity guarantee (SWHID)
  - Traceability (SWH Graph)

- And lot of knowledge exposed about open source



Radar

- On which projects my organization has contributed to ?

- Institutional portal for universities
- Software Heritage global archive can help
- Still need to design and implement it
- But also to improve our coverage and speed

Software Heritage

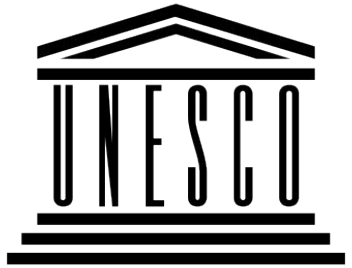
THE GREAT LIBRARY OF SOURCE CODE

You can help

- Becoming a sponsor
- Becoming a mirror
- Advocating for Software Heritage
- Joining the communities around Software Heritage
- Becoming an early user, helping us build tool tailored for your need

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE



United Nations
Educational, Scientific and
Cultural Organization

Thank you !

https://www.softwareheritage.org/
https://archive.softwareheritage.org/
https://gitlab.softwareheritage.org/
https://codecommons.org/
https://swhid.org/
https://swhsec.github.io/

thomas.aynaud@inria.fr

Paris Call

«Software source code represents unique knowledge of humanity's recent history.

It is therefore crucial to work together collectively so that the knowledge embedded in software source code is properly preserved, valued and sharedwith all.

This lies at the core of UNESCO's cooperation with Inria to support the creation of Software Heritage, the global archive of software source code»