

# Software Heritage

l'archive universelle du logiciel libre

David Douard

Engineers, Software Heritage – Inria

10 October 2024

Zenika

Nantes



# Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

# Software is all around us



# Software is built from *Source Code*

Harold Abelson, *Structure and Interpretation of Computer Programs* (1st ed.)

1985

*“Programs must be written for people to read, and only incidentally for machines to execute.”*

# Software is built from *Source Code*

Harold Abelson, *Structure and Interpretation of Computer Programs* (1st ed.)

1985

*“Programs must be written for people to read, and only incidentally for machines to execute.”*

## Apollo 11 source code (excerpt)

```
P63SPOT3      CA      BIT6      # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND   CHAN33
              EXTEND
              BZF    P63SPOT4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF    CODE500      # ASTRONAUT:  PLEASE CRANK THE
              TC     BANKCALL     #                SILLY THING AROUND
              CADR   GOPERF1
              TCF    GOTOP00H     # TERMINATE
              TCF    P63SPOT3     # PROCEED    SEE IF HE'S LYING

P63SPOT4      TC     BANKCALL     # ENTER      INITIALIZE LANDING RADAR
              CADR   SETPOS1

              TC     POSTJUMP     # OFF TO SEE THE WIZARD ...
              CADR   BURNBABY
```



# Software is built from *Source Code*

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.)

1985

*“Programs must be written for people to read, and only incidentally for machines to execute.”*

## Apollo 11 source code (excerpt)

```
P63SPOT3      CA      BIT6      # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND     CHAN33
              EXTEND
              BZF      P63SPOT4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF      CODE500      # ASTRONAUT:   PLEASE CRANK THE
              TC       BANKCALL     #             SILLY THING AROUND
              CADR     GOPERF1
              TCF      GOTOPOOH     # TERMINATE
              TCF      P63SPOT3     # PROCEED    SEE IF HE'S LYING

P63SPOT4      TC       BANKCALL     # ENTER      INITIALIZE LANDING RADAR
              CADR     SETPOS1

              TC       POSTJUMP     # OFF TO SEE THE WIZARD ...
              CADR     BURNBABY
```

## Parcoursup source code (excerpt)

```
public class AlgoOrdreAppel {

    /* la boucle principale de calcul des ordres d'appels.
       Renvoie une exception en cas de problème. */
    public static AlgoOrdreAppelSortie calculerOrdresAppels(AlgoOrdreAppelEntree data) throws VerificationException {

        VerificationEntreeAlgoOrdreAppel.verifier(data);

        AlgoOrdreAppelSortie resultat = new AlgoOrdreAppelSortie();
        /* calcul de l'ordre d'appel de chaque groupe de classement */
        for (GroupeClassement ga : data.groupeClassements) {
            resultat.ordresAppel.put(ga.cgpCod, ga.calculerOrdreAppel());
        }

        /* vérification avant retour des résultats */
        new VerificationsResultatsAlgoOrdreAppel().verifier(data, resultat);

        return resultat;
    }

    private AlgoOrdreAppel() {
    }
}
```

# Software is built from *Source Code*

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.)

1985

*“Programs must be written for people to read, and only incidentally for machines to execute.”*

## Apollo 11 source code (excerpt)

```
P63SP0T3      CA      BIT6      # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND    CHAN33
              EXTEND
              BZF     P63SP0T4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF     CODE500      # ASTRONAUT:   PLEASE CRANK THE
              TC      BANKCALL     #             SILLY THING AROUND
              CADR    GOPERF1
              TCF     G0T0P00H     # TERMINATE
              TCF     P63SP0T3     # PROCEED    SEE IF HE'S LYING

P63SP0T4      TC      BANKCALL     # ENTER      INITIALIZE LANDING RADAR
              CADR    SETPOS1

              TC      POSTJUMP     # OFF TO SEE THE WIZARD ...
              CADR    BURNBABY
```

## Parcoursup source code ( excerpt )

```
public class AlgoOrdreAppel {

    /* la boucle principale de calcul des ordres d'appels.
       Renvoie une exception en cas de problème. */
    public static AlgoOrdreAppelSortie calculerOrdresAppels(AlgoOrdreAppelEntree data) throws VerificationException {

        VerificationEntreeAlgoOrdreAppel.verifier(data);

        AlgoOrdreAppelSortie resultat = new AlgoOrdreAppelSortie();
        /* calcul de l'ordre d'appel de chaque groupe de classement */
        for (GroupeClassement ga : data.groupeClassements) {
            resultat.ordresAppel.put(ga.cgpCod, ga.calculerOrdreAppel());
        }

        /* vérification avant retour des résultats */
        new VerificationsResultatsAlgoOrdreAppel().verifier(data, resultat);

        return resultat;
    }

    private AlgoOrdreAppel() {
    }
}
```

Len Shustek, Computer History Museum

2006

*“Source code provides a view into the mind of the designer.”*

# Software source code as a key asset of Humankind

Experts call for greater recognition of software source code as heritage for sustainable development

6 November 2018



UNESCO, Inria, Software Heritage invite  
40 international experts meet in Paris ...

# Software source code as a key asset of Humankind

Experts call for greater recognition of software source code as heritage for sustainable development

6 November 2018



UNESCO, Inria, Software Heritage invite  
40 international experts meet in Paris ...



The call is published on February 2019

# Software source code as a key asset of Humankind

Experts call for greater recognition of software source code as heritage for sustainable development

6 November 2018



UNESCO, Inria, Software Heritage invite  
40 international experts meet in Paris ...

\* *“Recognise software source code as a fundamental enabler in all aspects of human endeavour”*



The call is published on February 2019

# Software source code is fragile

## Endangered source code ...



A word cloud containing terms such as: damage, disaster, malicious, deletion, obsolete, attack, dependencies, aging, tear, media, dangling, wear, corruption, encryption, format, reference, and storage.

- *link rot*: projects are created, moved around, removed
- *data rot*: physical media with legacy software decay
- *platform consolidation* endangers repositories
  - 2015 Google Code and Gitorious.org shutdown: ~1M
  - 2019 Bitbucket mercurial phase out: ~250.000
  - 2022 GitLab.com: **remove inactive projects?**

# Software source code is fragile

## Endangered source code ...



- *link rot*: projects are created, moved around, removed
- *data rot*: physical media with legacy software decay
- *platform consolidation* endangers repositories
  - 2015 Google Code and Gitorious.org shutdown: ~1M
  - 2019 Bitbucket mercurial phase out: ~250.000
  - 2022 GitLab.com: **remove inactive projects?**

... is endangered knowledge!

broken links and missing pieces in the *web of knowledge* of humankind

# Software source code is fragile

## Endangered source code ...



- *link rot*: projects are created, moved around, removed
- *data rot*: physical media with legacy software decay
- *platform consolidation* endangers repositories
  - 2015 Google Code and Gitorious.org shutdown: ~1M
  - 2019 Bitbucket mercurial phase out: ~250.000
  - 2022 GitLab.com: **remove inactive projects?**

... is endangered knowledge!

broken links and missing pieces in the *web of knowledge* of humankind

Bottomline: we need a global, long term effort

to build a *universal archive* of *all software source code*  
make it *resilient*  
and make it *sustainable*



*Unveiled in 2016*



## Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

*Unveiled in 2016*



## Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog



find and reference all  
software source code

*Unveiled in 2016*



## Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

### Reference catalog



**find** and **reference** all  
software source code

### Universal archive



**preserve** and **share** all  
software source code

Unveiled in 2016



## Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

### Reference catalog



find and reference all software source code

### Universal archive



preserve and share all software source code

### Research infrastructure



enable analysis of all software source code

# Today: a *universal* software archive, as a shared infrastructure

One infrastructure  
open and shared

Cultural Heritage



Industry



Research



Public Administration



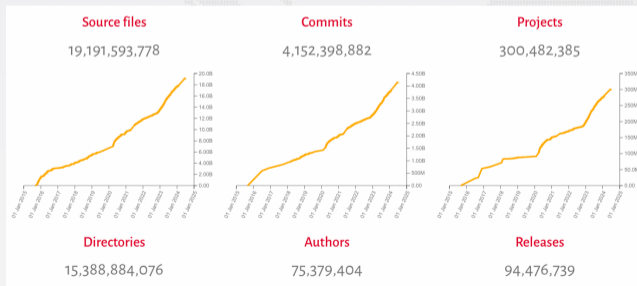
Software Heritage

# Today: a *universal* software archive, as a shared infrastructure

One infrastructure  
open and shared



The largest archive ever built



# Today: a *universal* software archive, as a shared infrastructure

One infrastructure  
open and shared

Cultural Heritage



Industry



Research

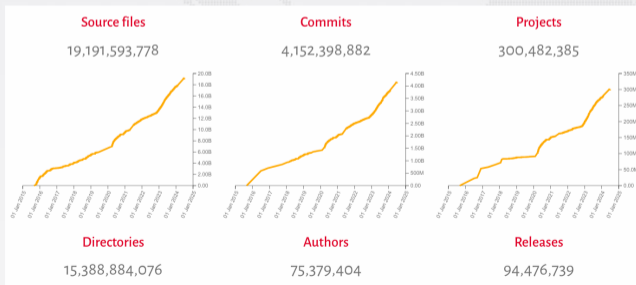


Public Administration



Software Heritage

The largest archive ever built

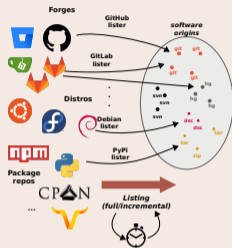


|                                |                           |                             |
|--------------------------------|---------------------------|-----------------------------|
| Bitbucket<br>2,509,402 origins | debian<br>136,338 origins | git<br>24,600 origins       |
| GitHub<br>197,883,004 origins  | gitlet<br>10,171 origins  | GitLab<br>4,216,298 origins |
| git<br>2,926 origins           | Gogs<br>172 origins       | GO<br>971,549 origins       |
| Guix<br>14,482 origins         | GNU<br>354 origins        | heptapod<br>1,207 origins   |
| launchpad<br>503,631 origins   | Maven<br>312,461 origins  | NixOS<br>14,482 origins     |

figures as of January 25 2024

# An operational, evolving infrastructure

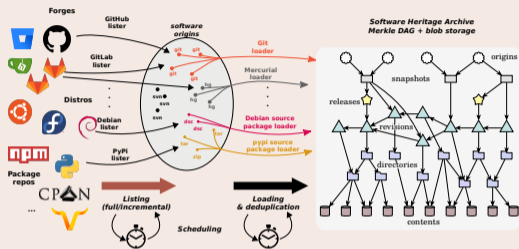
## Harvest and archive





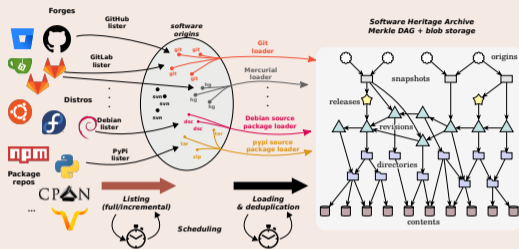
# An operational, evolving infrastructure

## Harvest and archive



# An operational, evolving infrastructure

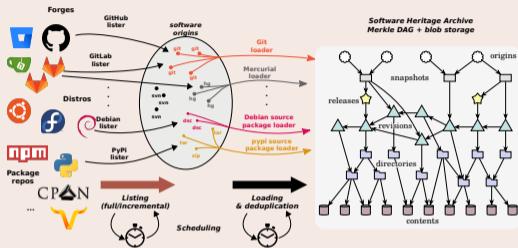
## Harvest and archive



- [save.softwareheritage.org](https://save.softwareheritage.org)
- [deposit.softwareheritage.org](https://deposit.softwareheritage.org)

# An operational, evolving infrastructure

## Harvest and archive



- [save.softwareheritage.org](https://save.softwareheritage.org)
- [deposit.softwareheritage.org](https://deposit.softwareheritage.org)

## Reference (35 billion SWHIDs)

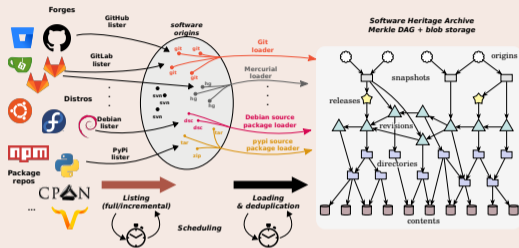
Intrinsic, decentralised, cryptographically strong identifiers



Now in SPDX 2.2, Wikidata, ISO is coming

# An operational, evolving infrastructure

## Harvest and archive



- [save.softwareheritage.org](https://save.softwareheritage.org)
- [deposit.softwareheritage.org](https://deposit.softwareheritage.org)

## Reference (35 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers



Now in SPDX 2.2, Wikidata, ISO is coming

Global development history permanently archived in a uniform data model

- over 20 billion unique source files from over 320 million software projects
- ~2PB (compressed) blobs, ~41 B nodes, ~650 B edges

Significant research challenges to explore it efficiently

# Archiving goals

Targets: VCS repositories & source code releases (e.g., tarballs, packages)

## We DO archive

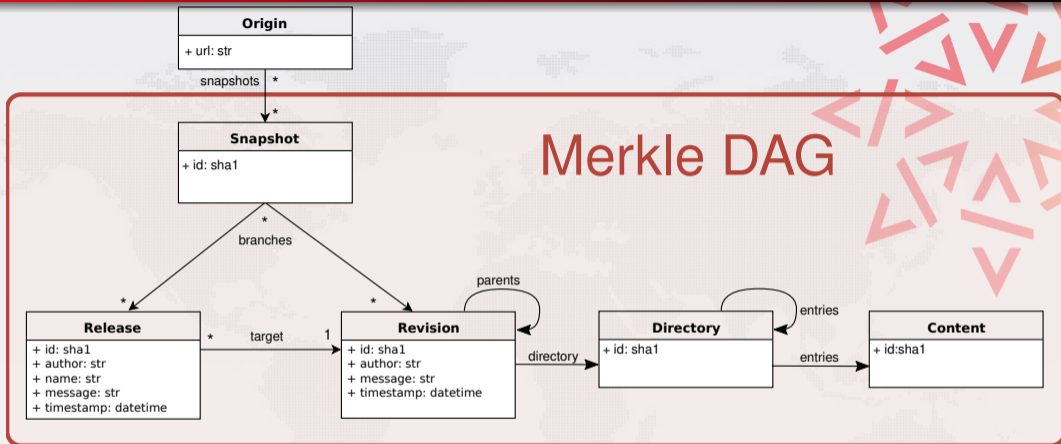
- file **content** (= blobs)
- **revisions** (= commits), with full metadata
- **releases** (= tags), ditto
- where (**origin**) & when (**visit**) we found any of the above

... in a VCS-/archive-agnostic **canonical data model**

## We DON'T archive (yet)

- homepages, wikis
- BTS/issues/code reviews/etc.
- mailing lists

Long term vision: play our part in a *"semantic wikipedia of software"*



A **global graph** linking together fully **deduplicated** source code artifact (files, commits, directories, releases, etc.) to the places that distribute them (e.g., Git repositories), providing a **unified view** on the entire *Software Commons*.

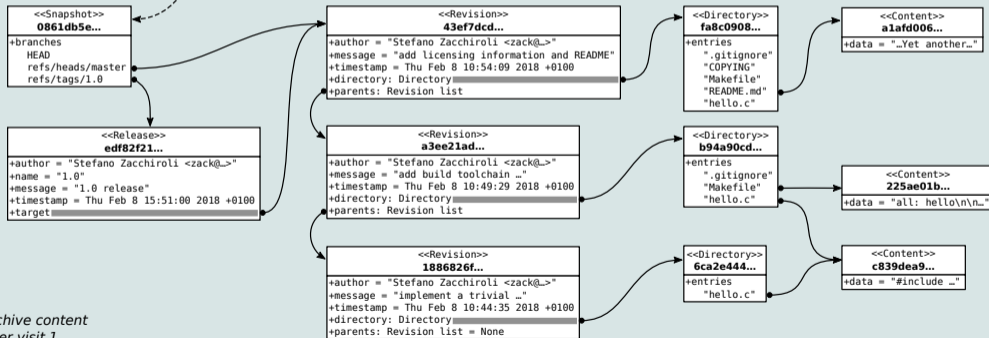
# The archive: a (giant) Merkle DAG

origin  
https://forge.softwareheritage.org/source/helloworld.git

visit  
1

snapshot  
0861db5e...

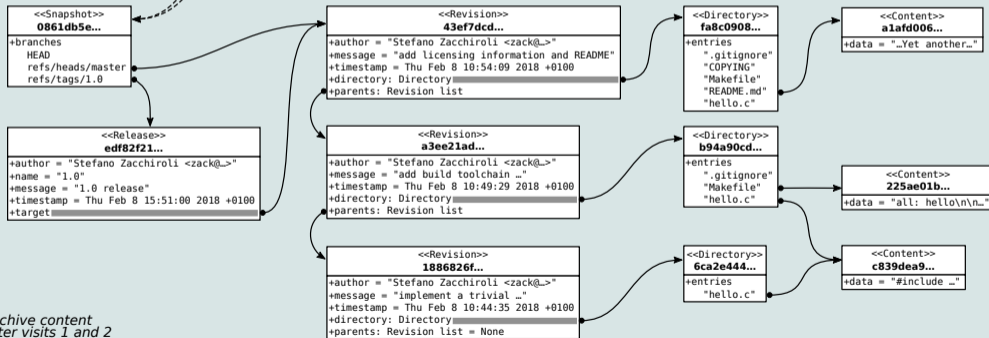
timestamp  
Fri Feb 9 12:38:45 2018 +0100



Archive content  
after visit 1

# The archive: a (giant) Merkle DAG

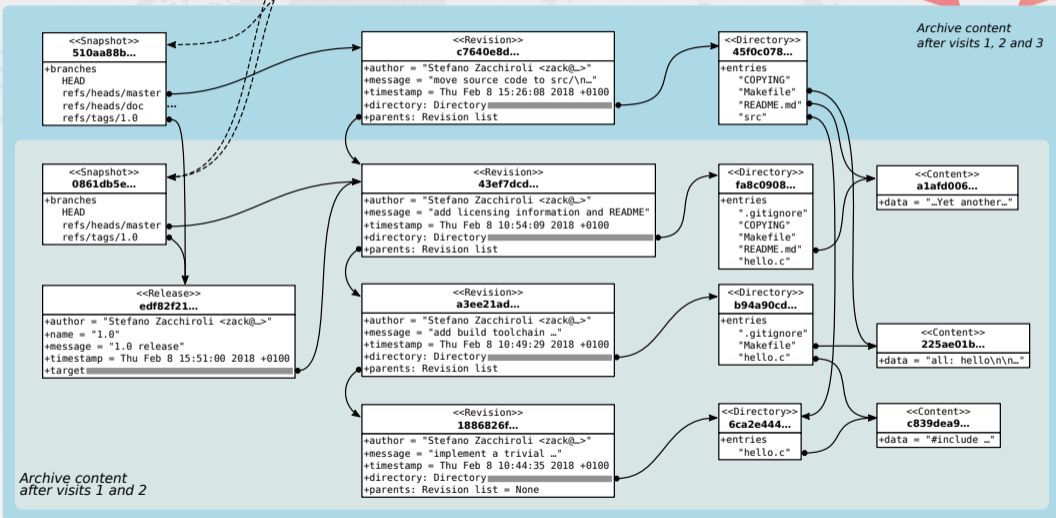
| origin   | visit | snapshot    | timestamp                     |
|--|-------|-------------|-------------------------------|
| https://forge.softwareheritage.org/source/helloworld.git | 1     | 0861db5e... | Fri Feb 9 12:38:45 2018 +0100 |
| https://forge.softwareheritage.org/source/helloworld.git | 2     | 0861db5e... | Fri Feb 9 13:29:00 2018 +0100 |





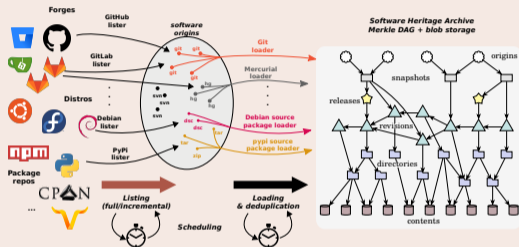
# The archive: a (giant) Merkle DAG

| origin   | visit | snapshot    | timestamp                     |
|--|-------|-------------|-------------------------------|
| https://forge.softwareheritage.org/source/helloworld.git | 1     | 0861db5e... | Fri Feb 9 12:38:45 2018 +0100 |
| https://forge.softwareheritage.org/source/helloworld.git | 2     | 0861db5e... | Fri Feb 9 13:29:00 2018 +0100 |
| https://forge.softwareheritage.org/source/helloworld.git | 3     | 510aa88b... | Fri Feb 9 15:52:50 2018 +0100 |



# An operational, evolving infrastructure

## Harvest and archive



- [save.softwareheritage.org](https://save.softwareheritage.org)
- [deposit.softwareheritage.org](https://deposit.softwareheritage.org)

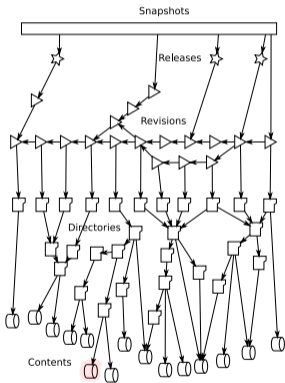
## Reference (40 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers



Now in SPDX 2.2, Wikidata, ISO standard

## Merkle DAG



## Content hash

### Contents

```
GNU GENERAL PUBLIC LICENSE
Version 3, 29 June 2007

Copyright (C) 2007 Free Software Foundation, Inc. <http://fsf.org/>
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

Preamble

The GNU General Public License is a free, copyleft license for
software and other kinds of works.

The licenses for most software and other practical works are designed
to take away your freedom to share and change the works. By contrast,
the GNU General Public License is intended to guarantee your freedom to
share and change all versions of a program--to make sure it remains free
software for all its users. We, the Free Software Foundation, use the
GNU General Public License for most of our software; it applies also to
any other work released this way by its authors. You can apply it to
your programs, too.

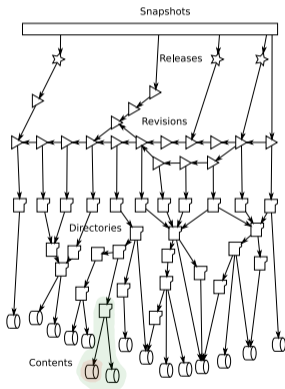
When we speak of free software, we are referring to freedom, not
price. Our General Public Licenses are designed to make sure that you
have the freedom to distribute copies of free software (and charge for
them if you wish), that you receive source code or can get it if you
want it, that you can change the software or use pieces of it in new
free programs, and that you know you can do these things.

To protect your rights, we need to prevent the license from being used to
```

sha1: 8624bcdae55baeef...  
sha256: 8ceb4b9ee5aded...  
sha1\_git: 94a9ed024d385...  
length: 35147

## Directory hash

### Merkle DAG



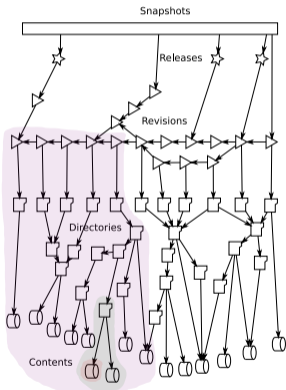
### Directories

- .gitignore
- AUTHORS
- LICENSE
- MANIFEST.in
- Makefile
- Makefile.local
- README.db\_testing
- README.dev
- bin
- debian
- docs
- requirements.txt
- setup.py
- sql
- swh
- utils

```
100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d0a34af6f52cf3cf6b0c2f7bd0648fbd255e77f AUTHORS
100644 blob 94a9ed024d3859793618152ea559a168bbcbb5e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e0867751dfb095c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5eaf2eb4 Makefile
100644 blob 5c7e3a5bbddb038682ba7793f440492ed9678bb3 Makefile.local
100644 blob 8617980629cd24e6080404f09aa749b085b3e07b README.db_testing
100644 blob 76b29f94cf815e0869c414d38d78d7ce08ec514e README.dev
040000 tree e1e10ecf948af0b93adb0372afc89f12e92618a bin
040000 tree 83e56d0beaf7793c77a45a345c80fcb8af503013 debian
040000 tree a34c9c4ba213f0cedc67f9816348d27955577af5 docs
100644 blob f2a6d32c6135aa7287bbd76167b01df2ae4f1539 requirements.txt
100755 blob eee147c36caf1bbc2d820da8dc026cb5b68180bc setup.py
040000 tree 224bb4c1f4c67fca1d160bffd2d06094e7e1abf3 sql
040000 tree 8631c9cd77bbe993168107ab5baf51f40c6300be swh
040000 tree 8fb905b56ba8ed692f1209b2773b474c6c1d66c1 utils
```


id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d

## Merkle DAG



## Revision hash

### Revisions

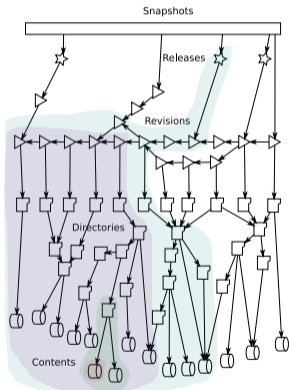
| Details  | Changes | Files |
|--|---------|-------|
| SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6  |         |       |
| Author: <a href="mailto:nicolas@dandrimont.eu">Nicolas Dandrimont &lt;nicolas@dandrimont.eu&gt;</a> (Thu Sep 1 14:26:13 2016)          |         |       |
| Committer: <a href="mailto:nicolas@dandrimont.eu">Nicolas Dandrimont &lt;nicolas@dandrimont.eu&gt;</a> (Thu Sep 1 14:26:13 2016)       |         |       |
| Subject: <code>provenance.tasks: add the revision -&gt; origin cache task</code>   |         |       |
| Parent: <a href="#">fc3a8b59ca1df424d860f2c29ab07fee4dc35d10</a> : test_storage: properly pipeline origin and cont...                  |         |       |
| provenance.tasks: add the revision -> origin cache task  |         |       |
| <a href="#">swh/storage/provenance/tasks.py</a>  77 |         |       |

tree [515f00d44e92c65322aaa9bf3fa097c00ddb9c7d](#)  
parent [fc3a8b59ca1df424d860f2c29ab07fee4dc35d10](#)  
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200  
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

provenance.tasks: add the revision -> origin cache task

id: [963634dca6ba5dc37e3ee426ba091092c267f9f6](#)

## Merkle DAG



## Release hash

### Releases

```
object c0c9f16b1e134f593e7567570a1761b156e6eb1d
type commit
tag v0.0.51
tagger Nicolas Dandrimont <nicolas@dandrimont.eu> 1472042163 +0200
```

Release swh.storage v0.0.51

- Add new metadata column to origin\_visit
- Update swh-add-directory script for updated API
- BEGIN PGP SIGNATURE—

```
tag v0.0.51
Tagger: Nicolas Dandrimont <nicolas@dandrimont.eu>
Date: Wed Aug 24 14:36:03 2016 +0200
```

Release swh.storage v0.0.51

- Add new metadata column to origin\_visit
- Update swh-add-directory script for updated API
- [...]

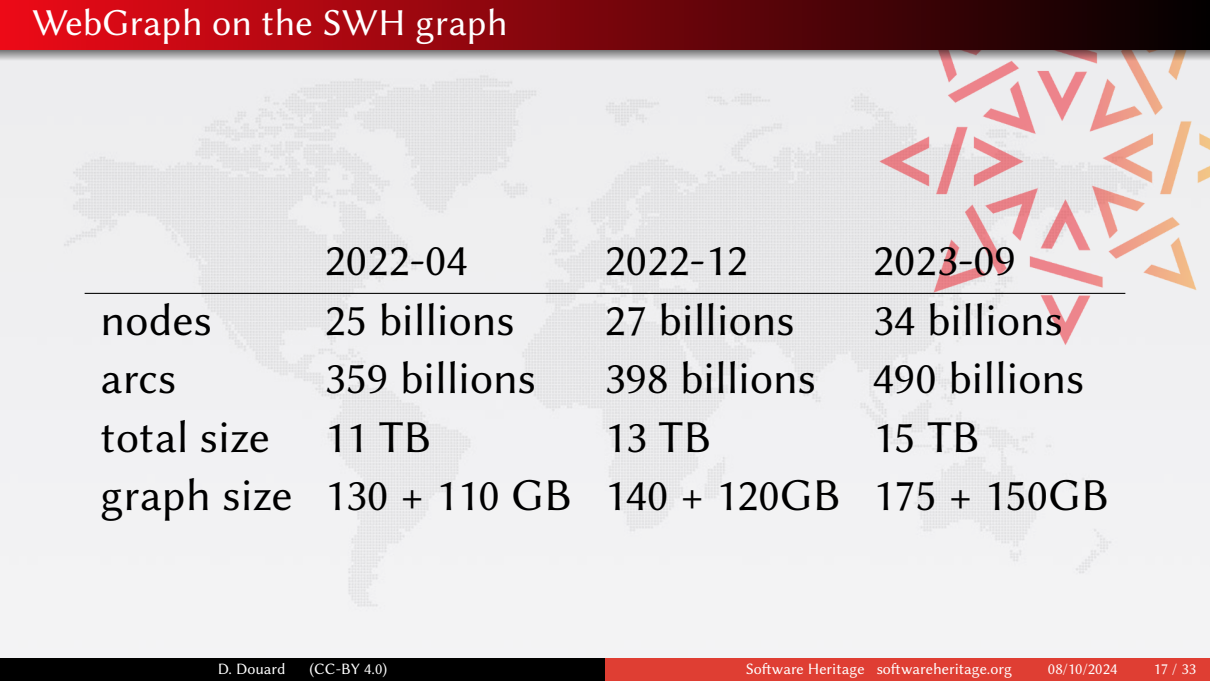
```
commit c0c9f16b1e134f593e7567570a1761b156e6eb1d
```

```
iQlzBAABCAAdBQJXvZTNFhxuaWNvbGFzQGRhbmRyaWV1bnQuXUuZACgkQ7AWLMO2+
neqor//aq6SOB5DijEa+kWN3rXgV5+1K1vEVh1wNKAwx8eKj7aX2kEILDt7uf
ahpZ6pz3q8nqs6aC1+YrxBfcih3L2YtrdZeWXWgr8xWNMaEoYDb8qaphwhBAD5t2
ICBlItZuItXuCrDt93eKkPwwzXGg+hB0sMwy35Dr6jW7Z7K4Mu/PgGlylHPY55yo
IGEndWno7VfH1Vm6t1n5qB7I5mXRaQA+becqddubtZ2xj+jiIUqC8cyqN3hm/IL
qsJ2mu8kyz3t8tG/H1/pV+I5OwBlnPoS5TH0tjogIEVgPK/dHSP79OuHDHFkCao
kjj6kAWyUB0Mxb+nKVjjeLbrR3+yWBFj3Qp5a1/V8oTh6E1dALcNmPEaKCoKtMt
d/gMRax1l1/g0EDfnsW67G6sDwKPKPhgtVLQ3nV3GaQQTnu1RpMz006H9tAwzC
Gg/K1PdH4HzOI46wYPZyje0U2VXGFu6vVU9vFQ4ZR/Wjn+0zMcZrdRlJSUOMn
RpTTU5sbXUeXHGOpgkXhSYTnvp1gdPc76U5TsK0aGe84AZm1Ik0mGrwXCVPqjYo
nhhibSBHBNMoqyF6yTSOpUbYK70tpYRRUGKWDerK0wK5xkWUzGtKzy6jYqJjo29
gulwgZQif5qWQCB0ontAL2+HvPfaVyckMejUhg62cP/+EHVUk=
=kOxP
—END PGP SIGNATURE—
```

id: [85083a5cc14a441c89dea73f5bdf67c3f9c6afdb](#)

- WebGraph: a framework to compress the graph of the web
- <https://webgraph.di.unimi.it/>
- Usually 1 to 3 bits per edge

# WebGraph on the SWH graph



|            | 2022-04      | 2022-12      | 2023-09      |
|------------|--------------|--------------|--------------|
| nodes      | 25 billions  | 27 billions  | 34 billions  |
| arcs       | 359 billions | 398 billions | 490 billions |
| total size | 11 TB        | 13 TB        | 15 TB        |
| graph size | 130 + 110 GB | 140 + 120GB  | 175 + 150GB  |



- Generic topology analysis
- Listing repositories containing a file/directory
- Finding the first revision containing a file/directory
- Listing the most common names for a given file
- Collaboration graph

## Inputs

- Intrinsic metadata: package.json, pom.xml, composer.json...
- Extrinsic metadata: proprietary formats

## Output

- JSON-LD (Schema.org/Codemeta and ForgeFed/ActivityStreams)

## Stream processors

- Based on Kafka

## Storage

- PostgreSQL (indexer storage)

## Indexing

- Elasticsearch

## Volumetry

- Actual archive: 270 million origins
- Indexed intrinsic metadata: 1.5 million origins
- Indexed extrinsic metadata: 60 million origins (mostly github)

## Mimetype (individual files)

- 14 billion rows in indexer storage
- one row per content

## Detected licences (using Fossology)

- 5.6 billion rows in indexer storage
- one row for each (content, license) pair

Search capabilities are currently limited to:

- origins (URLs)
- **some** metadata (MD)
  - intrinsic MD (known MD files translated to Codemeta)
  - extrinsic MD (for **some** forges only)

Uses elasticsearch

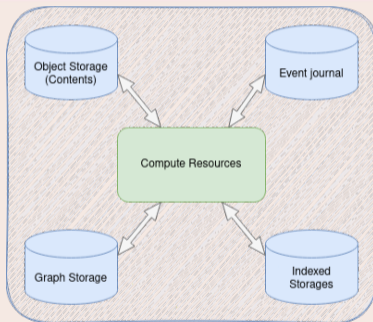
- index Origin URLs
- index Codemeta MD produced by the indexers

- full-text search on URLs (default behavior)
- full-text search on (Codemeta) intrinsic metadata
- also provides a simple Search Query Language
  - authenticated users only

## Examples:

- `origin : plasma and language in [python] and visits >= 5`
- `keyword in ["orchestration", "kubectl"] and license in ["GPLv3+", "GPLv3"]`

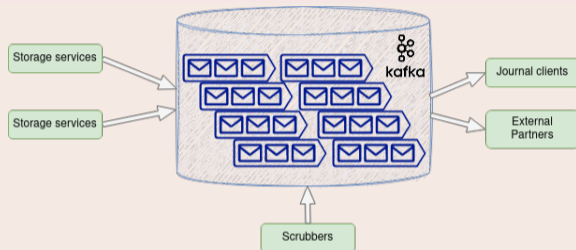
## Big picture



## In a few words...

- Mostly hosted in an INRIA's datacenter (Rocquencourt)
- 100% "commodity" hardware
- Object storage replicated on AWS and Azure object storages

## Kafka



## In a few words...

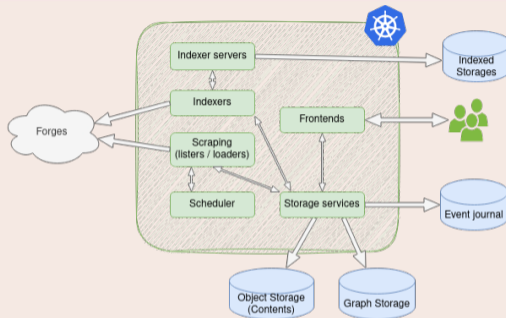
- Append only, full history of the archive
- 1 queue per type of object of the graph
- Used internally for event-driven architecture
- Used externally to keep mirrors in sync

## Hardware

- 4 servers
- 60TB of data
- 30TB of effective data



## Kubernetes



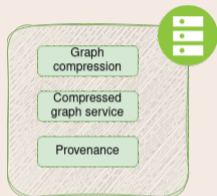
### In a few words...

- Kubernetes based
- Autoscaling of the services
- Orchestrated deployments

### Hardware

- 3 compute nodes + 1 management node
- Total: 336 cores / 768GB RAM

## Bare Metal



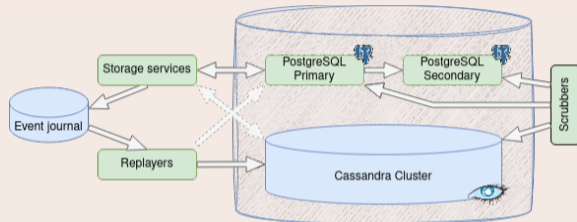
## In a few words...

- A couple of bare metal servers for specific needs
- A lot of cpus, disks and memory

## Hardware

- 48 cores / 521GB / 96To NVMe
- 48 cores / 4TB / 76To NVMe
- ...

## Storage backends



## In a few words...

- Monolithic PostgreSQL as main storage
- Migration to a distributed Cassandra backend in progress

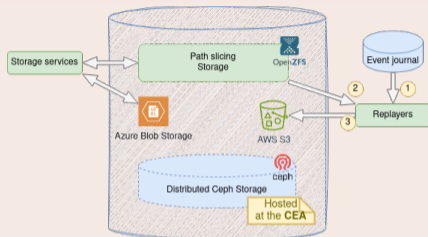
## Hardware - PostgreSQL

- 32 cores / 768GB
- Database: ~20To

## Hardware - Cassandra

- 10 servers
- ~85To of data
- Replication Factor: 3

## Storage backends



## In a few words...

- Monolithic main disk array storage
- Sync replication to Azure Blob Storage
- Async replication to Amazon S3
- Migration to a distributed storage in progress

## Hardware - Path slicing

- 17 Billion files
- 1.25PB compressed
- 1.9PB real

## Hardware - Ceph based

- ~2PB available w/ erasure coding
- Scalable architecture
  - 20 nodes of 144To
  - 2 management nodes
  - 2 index nodes

# A walkthrough

## General

- Browse [the archive](#), get and use SWHIDs, e.g. [Apollo 11 excerpt](#), [Parcoursup excerpt](#)
- [Trigger archival](#) with the [browser extension](#) or [webhook forge integration](#)

## Open Science

- [Curated deposit via HAL](#), e.g.: [LinBox](#), [SLALOM](#), [Givaro](#), [SumGra](#), [Coq proof](#), ...
- Cite software [with the biblatex-software style](#), e.g.: [article from IPOL](#)

## History of software: rescuing landmark legacy software

see [SWHAP process](#), [Software Stories](#), and [SWHAP Days 2022](#)

## Public code

Archived source code from [code.gouv.fr](#)

## Sharing the vision



United Nations  
Educational, Scientific and  
Cultural Organization



And many more ...

[www.softwareheritage.org/support/testimonials](http://www.softwareheritage.org/support/testimonials)

## Sharing the vision



United Nations  
Educational, Scientific and  
Cultural Organization



And many more ...

[www.softwareheritage.org/support/testimonials](http://www.softwareheritage.org/support/testimonials)

## Donors, members, sponsors

*Inria*

Diamond sponsor



Platinum sponsors



intel



Microsoft



Gold sponsors



openInventionnetwork

servicenow



SORBONNE  
UNIVERSITÉ

Université  
Paris Cité

Silver sponsors

AdaCore



GitHub

Google



Bronze sponsors



*we are all concerned, anyone can join and help*

# A growing and active community

## Core Team



## All together, 2024 Summit



## Ambassadors



Agustin Benito  
Bethencourt



Alexis Lebis



Anna-Lena  
Lamprecht



Bertrand Néron



Borut Kumperscak



Bostjan Spetic



Camille Françoise



Bruno Khelifi



Cécile Arènes



Dare Pejić



Flavia Marzano



Frédéric Santos



Gavin Henry



Gerard Coen



Gilmary Gallon



Harish Pillay



Italo Vignoli



Jaime Arias



Joëno Marques  
Da Costa



Julien Caugant



Malin Sandström



Maria-Chiara Prodi



Max Kalik



Maxence Azzouz-  
Thuëroz



Mohammad  
Akhlaghi



Neal Fultz



Ozbane Valencia



Pierre Poulain



Sandrine Layrise



Simon Phipps



Vicky Rampin



Violaine Louvet



Wendy Hagenmaier

[ambassadorprogram@softwareheritage.org](mailto:ambassadorprogram@softwareheritage.org)



# A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

*"a global infrastructure for open and better software at the service of humankind"*



## Software Heritage

[www.softwareheritage.org](http://www.softwareheritage.org)  
[@swheritage@mstdn.social](mailto:@swheritage@mstdn.social)

### Spread the word



- become an ambassador
- advocate for SWH in your communities