

# Software Heritage

l'archive universelle du logiciel libre

Nicolas Dandrimont & David Douard

Engineers, Software Heritage – Inria

16 March 2024

Assemblée Générale de l'April  
Sorbonne Université



# 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)

```
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
```

## 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

16 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 ...



The call is published on February 2019

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

# Software source code is fragile

## Endangered source code ...

A word cloud containing the following terms: damage, disaster, malicious, deletion, obsolete, attack, dependencies, aging, media, tear, dangling, wear, corruption, encryption, format, reference, and storage. The words are arranged in various sizes and orientations, with 'damage' and 'disaster' being the largest.

- *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



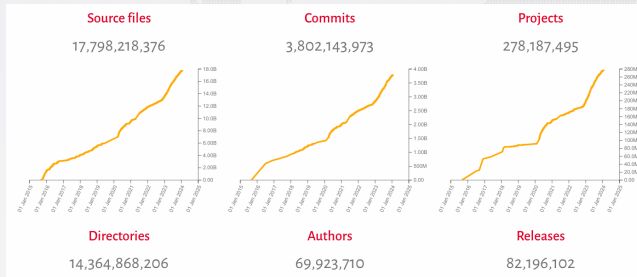
Software Heritage

# 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



The largest archive ever built

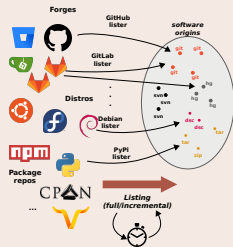


<b>Bitbucket</b> 2,509,402 origins	<b>debian</b> 136,338 origins	<b>git</b> 24,600 origins
<b>GitHub</b> 197,883,004 origins	<b>gitle</b> 10,171 origins	<b>GitLab</b> 4,216,298 origins
<b>git</b> 2,926 origins	<b>Gogs</b> 172 origins	<b>GO</b> 971,549 origins
<b>Guix</b> 14,482 origins	<b>GNU</b> 354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

figures as of January 25 2024

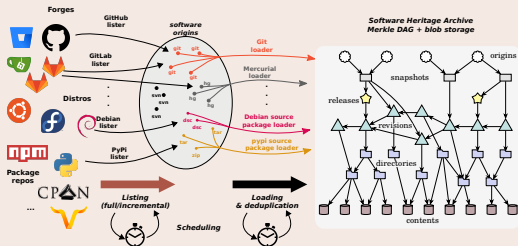
# An operational, evolving infrastructure

## Harvest and archive



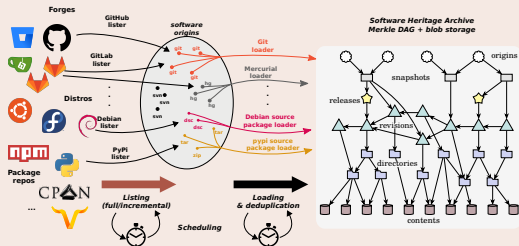
# An operational, evolving infrastructure

## Harvest and archive



# An operational, evolving infrastructure

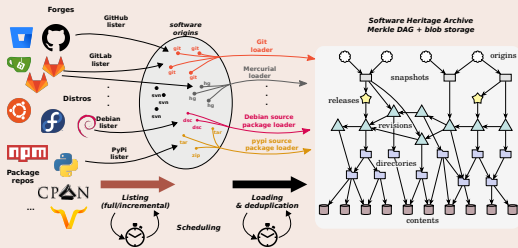
## Harvest and archive



- [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)

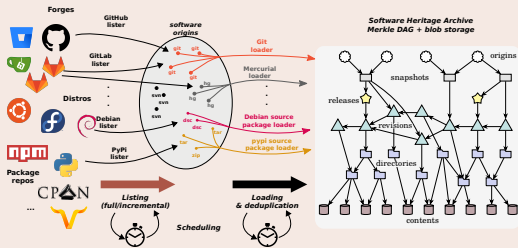
Intrinsic, decentralised, cryptographically strong identifiers



Now in SPDX 2.2, Wikidata, ISO is coming

# An operational, evolving infrastructure

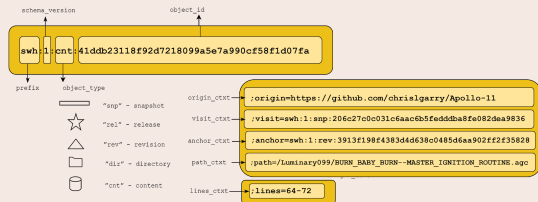
## Harvest and archive



- [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 17 billion unique source files from over 270 million software projects
- ~1.5PB (compressed) blobs, ~35 B nodes, ~500 B edges

Significant research challenges to explore it efficiently



# 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



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)

### We're hiring!



- sysadmin
- big data engineer
- backend developer

### Spread the word



- become an ambassador
- advocate for SWH in your communities