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 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.”

---

**Apollo 11 source code (excerpt)**

```
P63SPOT3
CA       BIT6       # IS THE LR ANTENNA IN POSITION 1 YET
EXTEND   RAND       CHAN33
EXTEND   BZC        P63SPOT4
              # BRANCH IF ANTENNA ALREADY IN POSITION 1
CAF       CODE509   # 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 {
    public static AlgoOrdreAppelSortie calculerOrdreAppels(AlgoOrdreAppelEntree data) throws VerificationException {
        VerificationEntreeAlgoOrdreAppel.verify(data);
        AlgoOrdreAppelSortie resultat = new AlgoOrdreAppelSortie();
        calculerOrdreAppelEntree pour chaque groupe de classement */
        for (GroupeClassement gs : data.groupesClassements) {
            resultat.ordreAppel.put(gs.classement, gs.calculerOrdreAppel());
        }
        VerificationResultatsAlgoOrdreAppel.verificationAvantRetourDesResultats(resultat);
        return resultat;
    }
    private AlgoOrdreAppel() {
    }
}
```

---

Len Shustek, *Computer History Museum* 2006

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

---

N. Dandrimont - D. Douard (CC-BY 4.0)
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 …

- **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*
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

Software Heritage

The largest archive ever built

**Source files**

17,798,218,376

**Commits**

3,802,143,973

**Projects**

278,187,495

**Directories**

14,364,868,206

**Authors**

69,923,710

**Releases**

82,196,102

**Bitbucket**

2,509,402 origins

**GitHub**

197,883,004 origins

**git**

24,400 origins

**debian**

56,983 origins

**gitlab**

26,599 origins

**Gogs**

136,338 origins

**Gnu**

10,171 origins

**Guix**

2,926 origins

**Heptapod**

4,216,290 origins

**launchpad**

172 origins

**Maven**

971,549 origins

**NixOS**

503,631 origins

**Npm**

312,461 origins

**Npm**

14,482 origins

**figures as of January 25 2024**

N. Dandrimont - D. Douard  (CC-BY 4.0)
An operational, evolving infrastructure

Harvest and archive

- save.softwareheritage.org
- deposit.softwareheritage.org

Reference (35 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers

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

N. Dandrimont - D. Douard (CC-BY 4.0)
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
An international, non-profit initiative for the long term

Sharing the vision

And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Diamond sponsor

Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors

we are all concerned, anyone can join and help

N. Dandrimont - D. Douard  (CC-BY 4.0)
A growing and active community

Core Team

Ambassadors

All together, 2024 Summit
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
@swheritage@mstdn.social

We’re hiring!
- sysadmin
- big data engineer
- backend developer

Spread the word
- become an ambassador
- advocate for SWH in your communities