Software Heritage

Large-scale Source Code Archival for Open Science

Stefano Zacchiroli

Télécom Paris — zack@upsilon.cc, @zacchiro

27 Jan 2022
IDIA, Institut Polytechnique de Paris

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

Software *source code* is precious knowledge

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
```

Quake III source code (excerpt)

```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    x2 = number * 0.5F;
    y = number;
    i = * ( long * ) &y; // evil floating point bit level hacking
    i = 0x5f3759df - ( i >> 1 ); // what the fuck?
    y = * ( float * ) &i;
    y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
// y = y * ( threehalfs - ( x2 * y * y ) ); // 2nd iteration, this
can be removed

    return y;
}
```

Len Shustek, Computer History Museum 2006

"Source code provides a view into the mind of the designer."

Calling for source code preservation: UNESCO

UNESCO, Inria, Software Heritage invite

40 international experts meet in Paris …

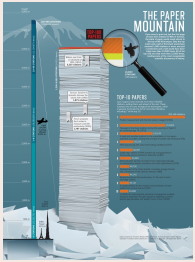
The call is published on Feb 2019

"[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive"

https://en.unesco.org/foss/paris-call-software-source-code

Source code history — for open science

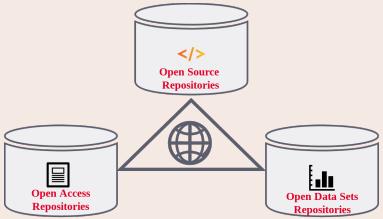
Software powers modern research



[…] software […] essential in their fields.
Top 100 papers (Nature, 2014)

Sometimes, if you dont have the software, you dont have the data
Christine Borgman, Paris, 2018

Missing pillar: software (source code)



The links in the picture are important

Nota Bene

software may be a *tool*, a *research outcome* and a *research objet*

access to the *source code* is essential!

Preserving the history of source code is important for *reproducibility*

Fragile

Like all digital information, FOSS is fragile

- link rot: projects are created, moved around, removed
- business-driven code loss (e.g., Gitorious, Google Code, Bitbucket)
- data rot: physical media with legacy software decay

If a website disappears you go to the Internet Archive…

where do you go if (a repository on) GitHub or GitLab goes away?

Outline

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 all software source code

Research infrastructure

enable analysis of all software source code

The largest public source code archive, principled

bit.ly/swhpaper

archive.softwareheritage.org

Technology

- transparency and FOSS
- replicas all the way down

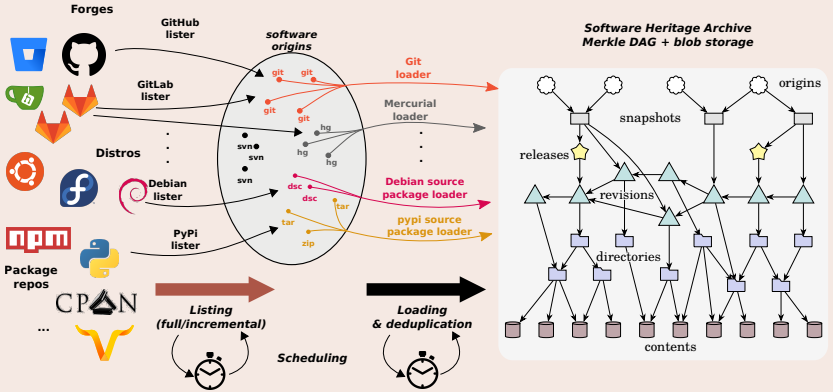
Content (billions!)

- intrinsic identifiers
- facts and provenance

Organization

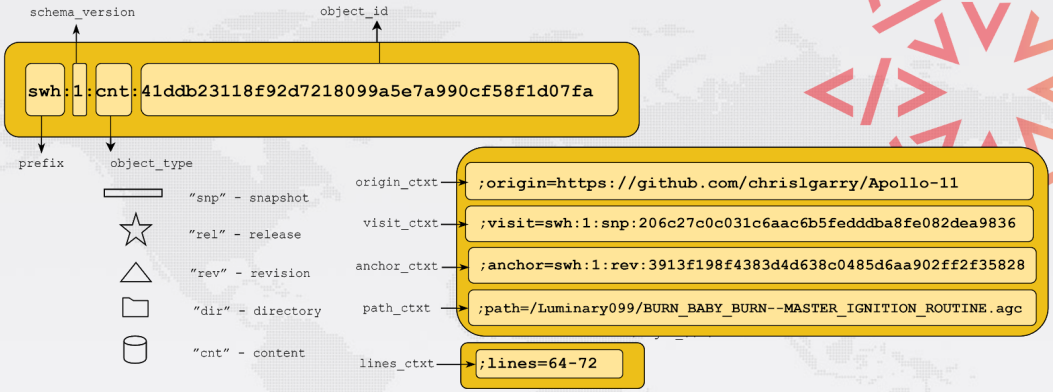
- non-profit
- multi-stakeholder

A peek under the hood: a global view on the software commons



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*.
(Size: ~20 B nodes, ~200 B edges, ~900 TB blobs)

Software Heritage *intrinsic* Identifiers (SWHID) (full spec)



An emerging standard

- in Linux Foundation's SPDX 2.2
- IANA registered, WikiData property P6138

Examples:

- Apollo 11 AGC excerpt
- Quake III rsqrt

Outline

1. Why we must preserve the history of software source code
2. How we can preserve our software heritage
3. Preserving our software commons: the present and the future

Focus on Academia: growing adoption (selection)

HAL software curated deposit workflow

Curated Archiving of Research Software Artifacts
International Journal of Digical Curation, 2020

Reference archive for swmath.org

swMATH
an information service for mathematical software

See *code* links, e.g. SemiPar package

IPOL (image processing)

- archive (deposit)
- reference
- BibLaTeX

eLife (life sciences)

- archive (save code now)
- reference

JTCAM (mechanics)

- instructions for authors
- biblatex-software in journal LaTeX class

Policy: France

NATIONAL PLAN FOR OPEN SCIENCE

National Plan for Open Science

Policy: Europe

EOSC SIRS report

- SWHIDs
- archive

Guidelines

- summary
- ICMS 2020

Sharing the vision

UNESCO
United Nations Educational, Scientific and Cultural Organization

And many more ...
www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Inria

Diamond sponsor
cea

Platinum sponsors
cnrs
HUAWEI
intel
Microsoft
SOCIETE GENERALE

Gold sponsors
openinventionnetwork
Université de Paris
SORBONNE UNIVERSITÉ

Silver sponsors
AdaCore
CAST
Google
GitHub
Università di Pisa
vmware

Bronze sponsors
DANS
FOSSID

Software Heritage and IDIA

Recommendations

- have an Open Science policy encompassing the trifecta (data, papers, source code)
- encourage researchers to:
  - archive source code used to support research work in Software Heritage (*GitHub is not an archive!*) → save.softwareheritage.org and/or HAL integration
  - reference source code from scientific papers using intrinsic, persistent identifiers → SWHID (Software Heritage IDentifiers) and biblatex-software
- guidelines for researchers: https://www.softwareheritage.org/save-and-reference-research-software/

Other ways of helping: engage with Software Heritage as an organization

- become a member/sponsor
- build a Software Heritage mirror