Software Heritage

an archive to enable our digital future

Roberto Di Cosmo

Director, Software Heritage
Inria and Université Paris Cité

February 1st 2024
UNESCO

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Software and Source Code

2. Software Heritage: a mission at the service of Humankind

3. Opening of the symposium

4. Symposium time

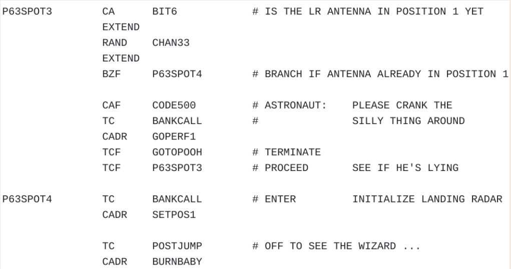
Software is all around us

Software is built from *Source Code*

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

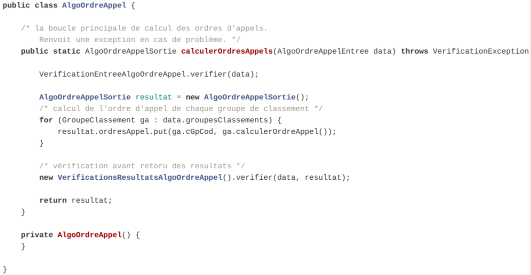
“Programs must be written for people to read, and only incidentally for machines to execute.”

Apollo 11 source code (excerpt)



```
P63SPOT3        CA      BIT6            # IS THE LR ANTENNA IN POSITION 1 YET
                EXTEND
                RAND    CHAN33
                EXTEND
                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)

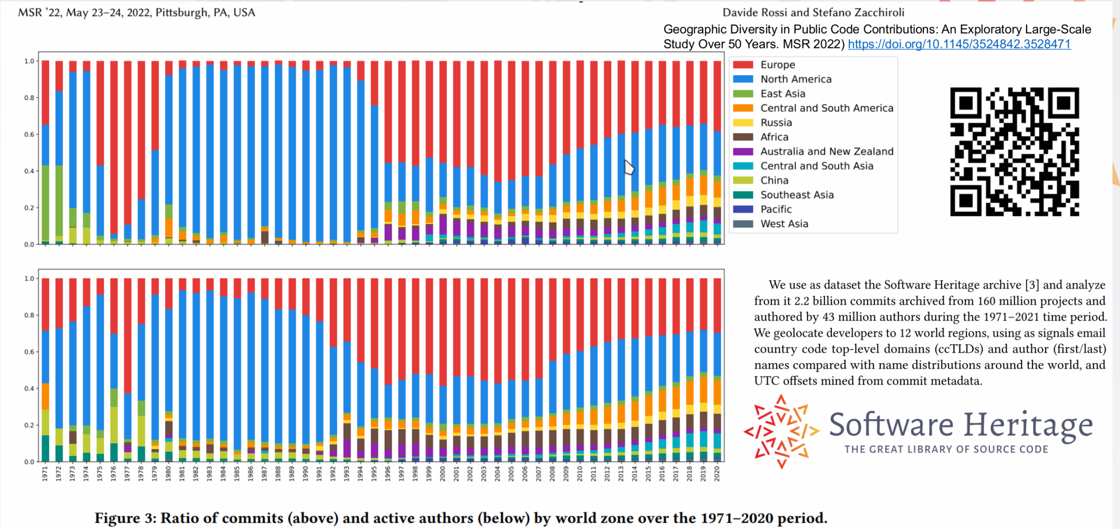


```java
public class AlgoOrdreAppel {

    /* la boucle principale de calcul des ordres d'appels.
        Renvoit 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.groupesClassements) {
            resultat.ordresAppel.put(ga.cGpCod, ga.calculerOrdreAppel());
        }

        /* vérification avant retoru des resultats */
        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.”

(Open) Source Code comes from all over the world



MSR '22, May 23–24, 2022, Pittsburgh, PA, USA

Davide Rossi and Stefano Zacchiroli

Geographic Diversity in Public Code Contributions: An Exploratory Large-Scale Study Over 50 Years. MSR 2022) https://doi.org/10.1145/3524842.3528471

We use as dataset the Software Heritage archive [3] and analyze from it 2.2 billion commits archived from 160 million projects and authored by 43 million authors during the 1971–2021 time period. We geolocate developers to 12 world regions, using as signals email country code top-level domains (ccTLDs) and author (first/last) names compared with name distributions around the world, and UTC offsets mined from commit metadata.

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Figure 3: Ratio of commits (above) and active authors (below) by world zone over the 1971–2020 period.

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

United Nations Educational, Scientific and Cultural Organization

Memory of the World

Inria inventors for the digital world

PARIS CALL

SOFTWARE SOURCE CODE

AS HERITAGE FOR SUSTAINABLE DEVELOPMENT

Software Heritage

UNESCO, Inria, Software Heritage invite

40 international experts meet in Paris …

The call is published on February 2019

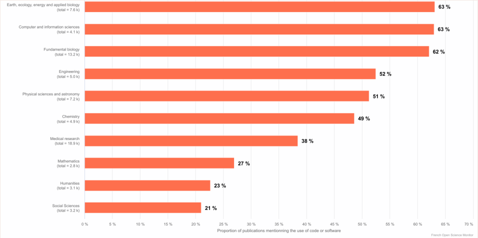
> *“Recognise software source code as a fundamental enabler in all aspects of human endeavour"*

(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

“The real antidote [to epidemic] is scientific knowledge *and* global cooperation.”

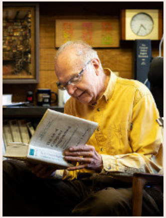
Software powers modern research



20%+ articles use software, all disciplines

2023 French Open Science Monitor

We can still talk to the early inventors



"Telling historical stories is the best way to teach. It’s much easier to understand something if you know the threads it is connected to."

Donald E. Knuth
Len Shustek
CACM, January 2021

We need a *dedicated infrastructure* to preserve and share *all* this knowledge!

Enhancing software Reuse, Security and Transparency

Software complexity is growing… it is important to Know Your SoftWare (KYSW)

Regulation on Software Updates

Recording […] software versions relevant to a vehicle type

UN Regulations on Cybersecurity, June 2020

Politique publique de la donnée, des algorithmes et des codes sources

…animer les ecosystèmes des…réutilisateurs du source code

Circulaire du Premier Ministre, 27 Avril 2021, France

Sec. 4. Enhancing Software Supply Chain Security

ensuring and attesting, to the extent practicable, to the integrity and provenance of open source software

May 2021 POTUS Executive Order

We need a *trusted* knowledge base with *software integrity and provenance* !

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*

Outline

1. Software and Source Code
2. Software Heritage: a mission at the service of Humankind
3. Opening of the symposium
4. Symposium time

Software Heritage in a nutshell

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

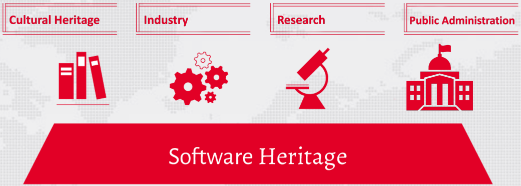
preserve and share all software source code

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

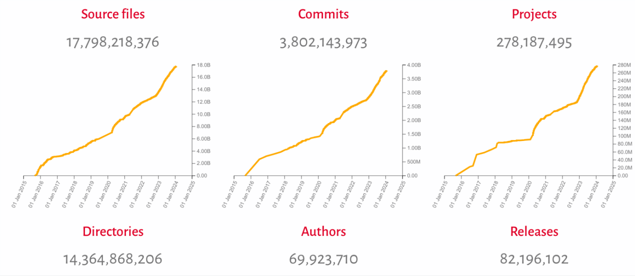
Industry

Research

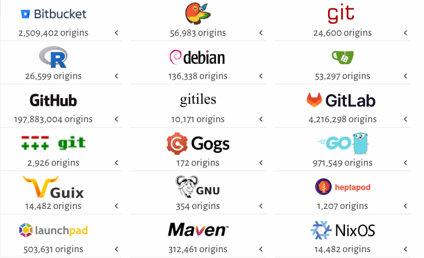
Public Administration

Software Heritage

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

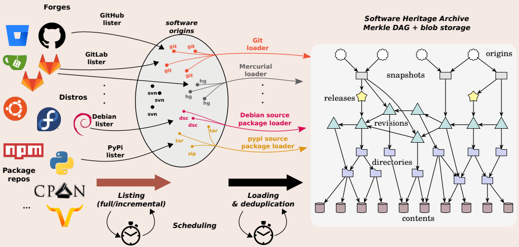


Bitbucket 2,509,402 origins	56,983 origins	git 24,600 origins
R 26,599 origins	debian 136,338 origins	53,297 origins
GitHub 197,883,004 origins	gitiles 10,171 origins	GitLab 4,216,298 origins
git 2,926 origins	Gogs 172 origins	GO 971,549 origins
Guix 14,482 origins	GNU 354 origins	heptapod 1,207 origins
launchpad 503,631 origins	Maven 312,461 origins	NixOS 14,482 origins

figures as of January 25 2024

An operational, evolving infrastructure

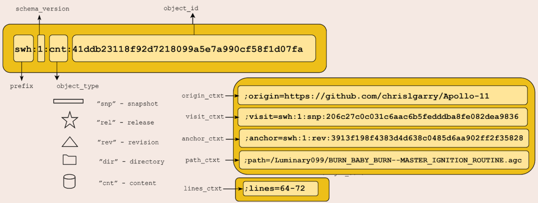
Harvest and archive



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

Reference (35 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers



```
swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa
;origin=https://github.com/chrislgarry/Apollo-11
;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836
;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828
;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc
;lines=64-72
```

Now in SPDX 2.2, Wikidata, ISO is coming

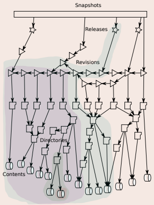
Global development history permanently archived in a uniform data model

- 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 (more later today)

A revolutionary infrastructure

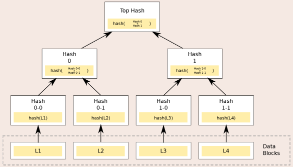
The *graph* of public software development



All software development in a single graph …

- enable traceability

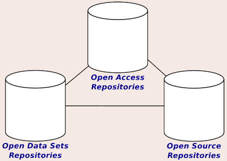
The *global ledger* of public code



… a Merkle graph

- ensure integrity

A *pillar* of Open Science

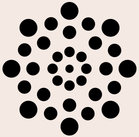


Reference archive of Research Software

- reproducibility
- reference

Reference platform for *Big Code*

uniform data structure



- large scale studies
- machine learning, AI, …

more later today

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

UNESCO
United Nations
Educational, Scientific and
Cultural Organization

IEEE Software · GitLab · ENGINEERING · eclipse
OW2 · fsfe · nlnet FOUNDATION · SIGSOFT · Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE · INAPI
THE LINUX FOUNDATION
ADULLACT · INFORMATICS EUROPE · Computer History Museum · software freedom conservancy
Software Freedom Law Center · AdaCore The GNAT Pro Company · gandi.net · TDF
FREE SOFTWARE FOUNDATION
creative commons · SIF Société Informatique de France · open source initiative · openinventionnetwork

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

Donors, members, sponsors

inria

Diamond sponsor

cea

Platinum sponsors

cnrs · SOCIETE GENERALE · intel
Microsoft · MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE · HUAWEI

Gold sponsors

Hugging Face · openinventionnetwork · servicenow · SORBONNE UNIVERSITÉ · Université de Paris

Silver sponsors

AdaCore · MINISTÈRE DES ARMÉES · DGA · RÉPUBLIQUE FRANÇAISE · GitHub · Google · UNIVERSITÀ DI PISA

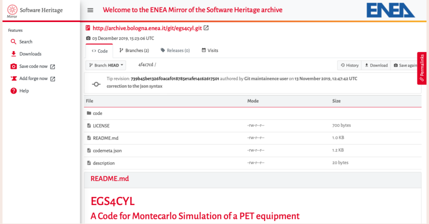
Bronze sponsors

Red Hat · SCANOSS · SCUOLA NORMALE SUPERIORE

we are all concerned, anyone can join and help

2023 progress hilights: preservation, recognition and AI

First international mirror at ENEA



Welcome to the ENEA Mirror of the Software Heritage archive

http://archive.bologna.enea.it/git/egs4cyl.git

03 December 2019, 15:23:06 UTC

4f4c7cd /

Tip revision: 739b45be1326f0acaf018785e1afe14c62617501 authored by Git maintainence user on 13 November 2019, 12:47:42 UTC
correction to the json syntax

File	Mode	Size
code		
LICENSE	-rw-r--r--	700 bytes
README.md	-rw-r--r--	1.0 KB
codemeta.json	-rw-r--r--	1.2 KB
description	-rw-r--r--	20 bytes

README.md

EGS4CYL

A Code for Montecarlo Simulation of a PET equipment

Opening at ENEA in Rome, 13/12/2023

Recognition as Open Science service

Software Heritage selected for 2024-2027

Principles for generative AI 10/2023

Open model

Data transparency

Author respect

… and much more

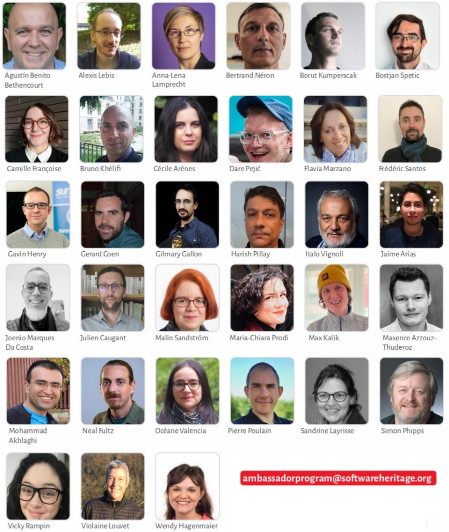
2023 annual report is here →

A growing and active community

Core Team

All together, 2023 Symposium

Ambassadors



Agustín Benito Bethencourt
Alexis Lebis
Anna-Lena Lamprecht
Bertrand Néron
Borut Kumperscak
Bostjan Spetic
Camille Françoise
Bruno Khélifi
Cécile Arènes
Dare Pejić
Flavia Marzano
Frédéric Santos
Gavin Henry
Gerard Coen
Gilmary Gallon
Harish Pillay
Italo Vignoli
Jaime Arias
Joenio Marques Da Costa
Julien Caugant
Malin Sandström
Maria-Chiara Prodi
Max Kalik
Maxence Azzouz-Thuderoz
Mohammad Akhlaghi
Neal Fultz
Océane Valencia
Pierre Poulain
Sandrine Layrisse
Simon Phipps
Vicky Rampin
Violaine Louvet
Wendy Hagenmaier

ambassadorprogram@softwareheritage.org

Outline

1. Software and Source Code
2. Software Heritage: a mission at the service of Humankind
3. Opening of the symposium
4. Symposium time

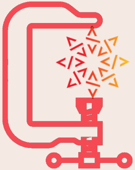
Today's main topics

from the connecting hub

Industry and Governments panel

- Digital tranformation
- Compliance and security
- Code as digital commons
- Open source and the SDG

Analyzing and Learning from the Archive



Presentations

- Fitting the SWH graph in main memory
- Building LLMs for code

Open Science panel

- Policy
- Infrastructures
- Funding

Cultural Heritage and Commons panel

Panel

- Memory of the World
- History of Software
- Digital Commons

Outline

1. Software and Source Code

2. Software Heritage: a mission at the service of Humankind

3. Opening of the symposium

4. Symposium time