Software Heritage
an archive to enable our digital future

Roberto Di Cosmo
Director, Software Heritage
Inria and Université Paris Cité

February 7th 2023
UNESCO
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)**

```assembly
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 BANCKALL # SILLY THING AROUND
   CADR GOPERF1
   TCF GOTOPO0H # TERMINATE
   TCF P63SP0T3 # PROCEED SEE IF HE’S LYING
P63SP0T4 TC BANCKALL # ENTER INITIALIZE LANDING RADAR
   CADR SETP0S1
   TC POSTJUMP # OFF TO SEE THE WIZARD ...
```

**Parcoursup source code (excerpt)**

```java
public class AlgoOrdreAppel {
    /* la boucle principale de calcul des ordres d'appels.
       Renvoie une exception en cas de problème. */
    public static AlgoOrdreAppelSortie calculerOrdreAppels(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.ordreAppel.put(ga.clgCod, 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.”
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"
Yuval Noah Harari (on COVID 19)

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

Software powers modern scientific research

The top 100 papers

 [...] the vast majority describe experimental methods or software that have become essential in their fields.

Nature, October 2014

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 écosystèmes 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 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 and make it sustainable
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

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**
13,841,373,660

**Commits**
2,886,851,936

**Projects**
204,290,404

**Directories**
11,265,049,858

**Authors**
53,633,976

**Releases**
36,535,307

---

- Bitbucket: 2,012,133 origins
- git: 19,494 origins
- GitHub: 21,486 origins
- Debian: 179,217 origins
- gitlab: 6,424 origins
- GitLab: 153,282,093 origins
- Guix: 12,451 origins
- Git: 354 origins
- Heptapod: 1,096 origins
- Guix: 356,873 origins
- Maven: 93,710 origins
- NixOS: 12,451 origins
- Maven: 1,799,296 origins
- Phabricator: 185 origins
- Python: 427,135 origins
- SourceForge: 308,970 origins

---

R. Di Cosmo  roberto@dicosmo.org  @rdicosmo  (CC-BY 4.0)
An operational, evolving infrastructure

Harvest and archive

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

Reference (25 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers

Now in SPDX 2.2, Wikidata, ISO is coming
A revolutionary infrastructure

The *graph* of Software Development

All software development in a single graph ...

- enable traceability

The *blockchain* of Software Development

... 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, ...

---

R. Di Cosmo  roberto@dicosmo.org  @rdicosmo  (CC-BY 4.0)
A walkthrough

General
- Browse the archive, get and use SWHIDs, e.g. Apollo 11 excerpt, Parcoursup excerpt
- Trigger archival in one click with the browser extension

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

Donors, members, sponsors

And many more ...

www.softwareheritage.org/support/testimonials

R. Di Cosmo  roberto@dicosmo.org  @rdicosmo  (CC-BY 4.0)
A growing and active community

Team

MIRRORS AND STORAGE PARTNERS

“Let us save what remains: … by such a multiplication of copies, as shall place them beyond the reach of accident.”

— Thomas Jefferson

Enea, GRNET, … CEA, RedHat

Ambassadors

All together, 5 years celebration in 2021

R. Di Cosmo roberto@dicosmo.org @rdicosmo (CC-BY 4.0)
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

Cultural Heritage and Education
Panel
- Memory of the World
- Digital preservation
- Software literacy

Long term source code preservation
Presentations
- 10,000 years on DNA
- shrinking the archive
- the first mirror

Open Science
Panel
- A view from NASA
- Supporting Open Science in Europe
- Words from UNESCO

Industry and Public administration
Presentations
(Open Source) Software
- Intel and beyond
- Foundations
- EU DIGIT OSPO
- Innovation in France