Software Heritage
an archive to enable our digital future

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

February 1st 2024
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)

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P63SP0T3</td>
<td># IS THE LR ANTENA IN POSITION 1 YET</td>
</tr>
<tr>
<td>EXTEND</td>
<td>CA</td>
</tr>
<tr>
<td>RAND</td>
<td>CHAN33</td>
</tr>
<tr>
<td>EXTEND</td>
<td>BZF</td>
</tr>
<tr>
<td>CAF</td>
<td>CODE508</td>
</tr>
<tr>
<td>TC</td>
<td>BANCKALL</td>
</tr>
<tr>
<td>CADR</td>
<td>GOPERF1</td>
</tr>
<tr>
<td>TCF</td>
<td>GOTOPOOH</td>
</tr>
<tr>
<td>TCF</td>
<td>P63SP0T3</td>
</tr>
<tr>
<td>P63SP0T4</td>
<td>TC</td>
</tr>
<tr>
<td>CADR</td>
<td>SETPOS1</td>
</tr>
<tr>
<td>TC</td>
<td>POSTJUMP</td>
</tr>
<tr>
<td>CADR</td>
<td>BURNBABY</td>
</tr>
</tbody>
</table>

Parcoursup source code (excerpt)

```java
public class AlgoOrdreAppel {
    /* la boucle principale de calcul des ordres d'appels. 
    Renvoie une exception en cas de probleme. */
    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.classement, ga.calculerOrdreAppel());
        }
        /* verification avant retour des resultats */
        new VerificationsRезультатAlgoOrdreAppel().Verifier(data, resultat);
        return resultat;
    }

    private AlgoOrdreAppel() {
    }
}
```

Len Shustek, Computer History Museum 2006

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

R. Di Cosmo roberto@dicosmo.org @rdicosmo (CC-BY 4.0)
We use the Software Heritage archive [3] and analyze it to 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.

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

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 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*
1. Software and Source Code

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

3. Opening of the symposium

4. Symposium time
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

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**
- 2,926 origins

**Guix**
- 14,482 origins

**launchpad**
- 503,631 origins

**Maven**
- 312,461 origins

**NixOS**
- 14,482 origins

**debian**
- 56,983 origins

**Gogs**
- 172 origins

**GNU**
- 354 origins

**GitLab**
- 4,216,296 origins

**heptapod**
- 1,207 origins

**Go**
- 4,216,296 origins

**heptapod**
- 1,207 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

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

Donors, members, sponsors

And many more ...

www.softwareheritage.org/support/testimonials

we are all concerned, anyone can join and help
2023 progress highlights: preservation, recognition and AI

First international mirror at ENEA
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
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

Industry and Governments panel
- Digital transformation
- 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