Software Heritage
saving the source code that powers our world

Roberto Di Cosmo
Director, Software Heritage

Nov 9th, 2019
Outline

1. Introduction
2. Saving our software heritage
Chi sono

Professore ordinario di Informatica, Parigi

(SNS 1986, Erdos #: 3)

- 30 anni di ricerca e insegnamento
- 20 anni di contribuzione al Software Libero
- 10 anni di direzione di strutture di interesse pubblico

1998 *Hold up planétaire* – best seller sui monopoli informatici
1999 *DemoLinux* – prima distribuzione live GNU/Linux
2007 *GTLL Systematic*  150 membri  40 progetti  200Me
2010 *IRILL* [www.irill.org](http://www.irill.org)
2015 *Software Heritage* [www.softwareheritage.org](http://www.softwareheritage.org)

@rdicosmo  Roberto Di Cosmo  @swheritage
So/f_tware is everywhere

Source code is **executable** and **human readable** knowledge

a growing part of our *Cultural Heritage*
"The source code for a work means the preferred form of the work for making modifications to it."

Hello World

Program (source code)
```
/* Hello World program */
#include<stdio.h>

void main()
{
    printf("Hello World");
}
```

Program (excerpt of binary)
```
<table>
<thead>
<tr>
<th>Address</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>4004e6</td>
<td>55</td>
</tr>
<tr>
<td>4004e7</td>
<td>48 89 e5</td>
</tr>
<tr>
<td>4004ea</td>
<td>bf 84 05 40 00</td>
</tr>
<tr>
<td>4004ef</td>
<td>b8 00 00 00 00</td>
</tr>
<tr>
<td>4004f4</td>
<td>e8 c7 fe ff ff</td>
</tr>
<tr>
<td>4004f9</td>
<td>90</td>
</tr>
<tr>
<td>4004fa</td>
<td>5d</td>
</tr>
<tr>
<td>4004fb</td>
<td>c3</td>
</tr>
</tbody>
</table>
```
Software source code: a precious part of our heritage

Harold Abelson, Structure and Interpretation of Computer Programs (1985)

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

Apollo 11 source code (excerpt)

```
P63SP03
CA BIT6  # IS THE LR ANTENNA IN POSITION 1 YET
EXTEND
RAND CHAN33
EXTEND
BZF P63SP04  # BRANCH IF ANTENNA ALREADY IN POSITION 1
CAF CODE500  # ASTRONAUT: PLEASE CRANK THE
TC BANKCALL  # SILLY THING AROUND
CADR GOPERP1
TCF GOTOPOOH  # TERMINATE
TCF P63SP03  # PROCEED SEE IF HE'S LYING
P63SP04
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 threehalves = 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 * ( threehalves - ( x2 * y * y ) ); // 1st iteration
    // y = y * ( threehalves - ( 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.”

@rdicosmo Roberto Di Cosmo @swheritage
... and only just half a century old!

Apollo 11 (~60,000 lines), 1969

"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."  
Margaret Hamilton

Linux Kernel (in your pockets!)

An exponential growth!

https://hal.archives-ouvertes.fr/hal-02158292v2
Free and Open Source software and society

Free Software: offers to its users the freedom to:

- use the software
- study and adapt the software
- distribute software copies
- distribute modified copies

Why bother?
Changes how software is developed, tested, maintained, marketed, sold

Why bother?
Gives back control to user / citizen, organization, enterprise, state

Algorithmic transparency

@rdicosmo  Roberto Di Cosmo  @swheritage
A precious growing commons

Three Layers emerging successively

1984 - 1998  focus on freedom for users and developers
1998 - 2014  focus on projects: quality, cost, community
2014 - ...  focus on governance, strategic alignment
... and an endangered heritage

We are at a turning point

key people are passing away …

software development skyrockets …

and yet we are missing …

A catalog

An archive

A research infrastructure

it is urgent to take action
Outline

1 Introduction

2 Saving our software heritage
Our mission

Collect, preserve and share the source code of all the software that is available

Past, present and future

Preserving the past, enhancing the present, preparing the future
An international, non profit initiative built for the long term

Sharing the vision

Donors, members, sponsors

And many more ...

www.softwareheritage.org/support/testimonials
Global development history permanently archived in a unique git-like Merkle DAG

- ~400 TB (uncompressed) blobs, ~20 B nodes, ~280 B edges
A revolutionary infrastructure for industry

The graph of Software Development

All of the software development in a single graph!
- lookup by content hash
- wayback machine for software development
  - http://archive.softwareheritage.org/
- ... and much more

The blockchain of Software Development

All of a software development... in a single Merkle graph!
Widely used crypto (e.g., Git, blockchains, IPFS, ...)
- built-in deduplication
- intrinsic, unforgeable identifiers at all levels
- simplifies traceability (licensing, supply chain management)
A revolutionary infrastructure for research and innovation

A pillar of Open Science

The *reference archive* of Research Software for Open Science
- curated deposit of research software
- in collaboration with HAL, CCSD and Inria IES
- now open *to all researchers!*
- intrinsic identifiers for reproducibility

Reference platform for *Big Code*
- unique *observatory* of all software development
- *big data, machine learning* paradise: classification, trends, coding patterns, code completion…
Milestones

Summer 2015
The collection starts: first server, (very) early prototype

June 30th 2016
Public unveiling, with the first sponsors: Microsoft and DANS

April 3rd 2017
Unesco - Inria agreement on software access and preservation.

June 7th 2018
Opening the archive to the world

December 7th 2018
Starting the mirror network

February 26th 2019
Publication of the expert meeting Paris Call on Software Source Code
Breaking news: archiving public code

https://code.etalab.gouv.fr
Breaking news: SWHAP

Paris Call on Software Source Code

“[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”

SWHAP: an important step forward

- detailed guidelines to curate landmark legacy source code and archive it on Software Heritage
- intense cooperation with Università di Pisa and UNESCO
- open to all, we’ll promote it worldwide

https://www.softwareheritage.org/swhap
Thomas Jefferson, February 18, 1791

…let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.

Welcoming ENEA

- first institutional mirror
- increased resilience
- AI infrastructure for researchers
- stepping stone to an European joint effort
Join the revolution!

www.softwareheritage.org @swheritage

Library of Alexandria/CERN of code

Working together to
- recover the past
- preserve our heritage
- share the knowledge
- prepare the future

Learn more www.softwareheritage.org/publications

- Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchirolı
  *Building the Universal Archive of Source Code*, Communications of the ACM, October 2018

- Roberto Di Cosmo, Morane Gruenpeter, Stefano Zacchirolı

- Roberto Di Cosmo, Stefano Zacchirolı