Software Heritage
Opening the doors to the public

Roberto Di Cosmo
roberto@dicosmo.org
June 7th, 2018
Software is everywhere
Software embodies a growing part of...

...our scientific, technical and Cultural Heritage!
“The source code for a work means the preferred form of the work for making modifications to it.”

GPL Licence
“The source code for a work means the preferred form of the work for making modifications to it.”

Hello World
"The source code for a work means the preferred form of the work for making modifications to it."

Hello World

Program (excerpt of binary)

4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
4004f9: 90
4004fa: 5d
4004fb: c3
Source Code: executable and human readable knowledge

“The source code for a work means the preferred form of the work for making modifications to it.”

Hello World

<table>
<thead>
<tr>
<th>Program (excerpt of binary)</th>
<th>Program (source code)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4004e6: 55</td>
<td>/* Hello World program */</td>
</tr>
<tr>
<td>4004e7: 48 89 e5</td>
<td>#include&lt;stdio.h&gt;</td>
</tr>
<tr>
<td>4004ea: bf 84 05 40 00</td>
<td>void main()</td>
</tr>
<tr>
<td>4004ef: b8 00 00 00 00</td>
<td>{</td>
</tr>
<tr>
<td>4004f4: e8 c7 fe ff ff</td>
<td>printf(&quot;Hello World&quot;);</td>
</tr>
<tr>
<td>4004f9: 90</td>
<td>}</td>
</tr>
<tr>
<td>4004fa: 5d</td>
<td></td>
</tr>
<tr>
<td>4004fb: c3</td>
<td></td>
</tr>
</tbody>
</table>
~ 50 years, a lightning fast growth

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

Margaret Hamilton

Apollo 11 (~60,000 lines), 1969

Linux Kernel (in your pockets!)
~ 50 years, a lightning fast growth

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!)

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

“Programs must be written for people to read, and only incidentally for machines to execute.”
~ 50 years, a lightning fast growth

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!)

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

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

Len Shustek, Computer History Museum (2006)

“Source code provides a view into the mind of the designer.”
We are at a turning point

Preserve the past

Only a few years left to recover the history of software technology
We are at a turning point

Preserve the past

Only a few years left to recover the history of software technology

Improve the future

We need a universal platform for all the future software source code
Collect, preserve and share the source code of all the software

Preserving our heritage, enabling better software and better science for all
Collect, preserve and share the source code of all the software

Preserving our heritage, enabling better software and better science for all

Reference catalog

find and reference all the source code
Software Heritage, in a nutshell

Collect, preserve and share the source code of all the software
Preserving our heritage, enabling better software and better science for all

Reference catalog
find and reference all the source code

Universal archive
preserve all the source code
Collect, preserve and share the source code of all the software

Preserving our heritage, enabling better software and better science for all

Reference catalog

find and reference all the source code

Universal archive

preserve all the source code

damage, disaster, malicious, deletion

Research infrastructure

enable analysis of all the source code

Roberto Di Cosmo

www.dicosmo.org
Milestones
Mai 2015: ignition

initial push

core team and infrastructure

long term commitment

open approach, and continued support
Our first press release
Raising awareness for Software Source Code preservation and access
Summer 2018: getting up to speed

The team

GitHub, Debian, Gitorious, GoogleCode, …

The archive

Sponsors

Microsoft  Intel  Société Générale
Huawei, Google
DANS, Nokia Bell Labs, University of Bologna,
GitHub, UQAM, FOSSID

Collaborations

- Qwant (search engine)
- Crossminer
- HAL, Adullact

Roberto Di Cosmo  www.dicosmo.org
Today: opening the doors!

Opening the door of the Software Heritage archive
Visiting the archive: the Apollo 11 source code

Margaret Hamilton today

The Apollo 11 source code in SWH

Some pointers

- Entry point
- Burn, baby, burn!
Visiting the archive: the Quake 3 source code

John Carmack

The Quake 3 source code in SWH

Some pointers

- Entry point
- What the f…
Yes, now you can!

Wayback machine for source code
- retrieve the source code as it was

Reference catalog
- use intrinsic identifiers for software

Open science
- deposit scientific software (via HAL)

Universal knowledge base
- store the knowledge about source code

And much, much more is in store!

see the strategic plan…
Everybody is concerned, everybody can help build

The Library of Alexandria of code
- recover the past
- structure the future

A CERN for Software
- build better software
  - for industry
  - for society as a whole