Software Heritage: global archive to enable our digital future

Roberto Di Cosmo

Director, Software Heritage
Inria and Université de Paris Cité

13 December 2023
Outline

1. Introduction
2. Software Heritage to the rescue
3. Library of Alexandria as an inspiration
4. The way forward
"The source code for a work means the preferred form of the work for making modifications to it."
“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)

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

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>

Program (source code)

```c
/* Hello World program */

#include<stdio.h>

void main()
{
    printf("Hello World");
}
```
“Programs must be written for people to read, and only incidentally for machines to execute.”


Software Source Code is Knowledge

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>P63SPOT3</th>
<th>CA</th>
<th>BIT6</th>
<th># IS THE LR ANTENNA IN POSITION 1 YET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EXTEND</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RAND</td>
<td>CHAN33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BZF</td>
<td>P63SPOT4</td>
<td># BRANCH IF ANTENNA ALREADY IN POSITION 1</td>
</tr>
<tr>
<td>CAF</td>
<td>CODE508</td>
<td># ASTRONAUT: PLEASE CRANK THE</td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>BANKCALL</td>
<td># SILLY THING AROUND</td>
<td></td>
</tr>
<tr>
<td>CADR</td>
<td>GOPERF1</td>
<td># TERMINATE</td>
<td></td>
</tr>
<tr>
<td>TCF</td>
<td>GOTO000H</td>
<td># PROCEED SEE IF HE'S LYING</td>
<td></td>
</tr>
<tr>
<td>TCF</td>
<td>P63SPOT3</td>
<td># PROCEED</td>
<td></td>
</tr>
</tbody>
</table>

| P63SPOT4 | TC  | BANKCALL | # ENTER INITIALIZE LANDING RADAR |
| CADR     | SETPOS1 | # OFF TO SEE THE WIZARD ... |
| TC       | POSTJUMP | BURNBABY |

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

Software Heritage as a key infrastructure   13/12/2023   2 / 14
“Programs must be written for people to read, and only incidentally for machines to execute.”

Apollo 11 source code (excerpt)

```
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 CODE508 # ASTRONAUT: PLEASE CRANK THE
   TC BANKCALL # SILLY THING AROUND
   CADR GOPERF1
   TCF GOTOP00H # TERMINATE
   TCF P63SP0T3 # PROCEED SEE IF HE’S LYING

P63SP0T4 TC BANKCALL # ENTER INITIALIZE LANDING RADAR
   CADR SETPOS1
   TC POSTJUMP # OFF TO SEE THE WIZARD ...
```

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;
}
```
Software Source Code is Knowledge

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 CODE508 # ASTRONAUT: PLEASE CRANK THE
   TC BANKCALL # SILLYThing AROUND
   CADR GOPERF1
   TCF GOTOP000H # 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 ...
```

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.”
Yuval Noah Harari (on COVID 19)

“The real antidote [to epidemic] is scientific knowledge and global cooperation.”
Yuval Noah Harari (on COVID 19)

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

Software powers modern scientific research

20+% articles across all disciplines share software

2023 French Open Science Monitor
Yuval Noah Harari (on COVID 19)

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

Software powers modern scientific research

20+% articles across all disciplines share software 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
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 scientific research

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

20+% articles across all disciplines share software 2023 French Open Science Monitor

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

2015 Google Code and Gitorious.org shutdown: ~1M endangered repositories

2019 250,000 endangered repositories on BitBucket
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*

  2015  Google Code and Gitorious.org shutdown:
  ~1M endangered repositories

  2019  250.000 endangered repositories on BitBucket

... is endangered knowledge!

broken links and missing pieces in the *web of knowledge* of humankind
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*

2015 Google Code and Gitorious.org shutdown:
- ~1M endangered repositories

2019 250,000 endangered repositories on BitBucket

… 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*
UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …
A global call for action

Experts call for greater recognition of software source code as heritage for sustainable development

UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …

The call is published on Feb 2019


R. Di Cosmo  roberto@dicosmo.org  (CC-BY 4.0)
A global call for action

UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …

“Support the development of shared infrastructures to collect, preserve and make available software source code; […] for the large scale analysis and improvement of the quality, safety and security of the software commons;”

Outline

1. Introduction

2. Software Heritage to the rescue

3. Library of Alexandria as an inspiration

4. The way forward
Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all
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
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
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
An international, non profit initiative built for the long term

Sharing the vision

And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Diamond sponsor

Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors
One infrastructure
open and shared

Software Heritage
The largest software archive, a shared infrastructure

One infrastructure
open and shared

Software Heritage

The largest archive ever built

Source files
16,786,947,392

Commits
3,555,307,942

Projects
262,077,475

Directories
13,486,743,792

Authors
63,835,635

Releases
74,477,682

R. Di Cosmo  roberto@dicosmo.org  (CC-BY 4.0)
The largest software archive, a shared infrastructure

One infrastructure open and shared

Software Heritage

The largest archive ever built

Source files: 16,786,947,392
Commits: 3,555,307,942
Projects: 262,077,475

Directories: 13,486,743,792
Authors: 63,835,635
Releases: 74,477,682

Bitbucket: 2,012,133 origins
git: 19,494 origins
GitHub: 21,486 origins
debian: 129,217 origins
Guix: 12,451 origins
gitlab: 3,089,638 origins
NixOS: 1,096 origins
launchpad: 356,873 origins
Maven: 93,710 origins
heptapod: 1,247 origins
npm: 1,799,396 origins
Phabricator: 185 origins
SourceForge: 398,970 origins
Outline

1. Introduction
2. Software Heritage to the rescue
3. Library of Alexandria as an inspiration
4. The way forward
The Great Library of Alexandria

Built by the Ptolemys (3rd C BC):
- send scholars all over the world…
- … to copy precious documents …
- … and bring them back under one roof!
Software Heritage under the hood

... to a modern Library of Alexandria for Software Source Code
… to a modern Library of Alexandria for Software Source Code

Software Heritage under the hood

Software Heritage by DALL-E

DALL-E’s view

Software Heritage as a key infrastructure 13/12/2023 10 / 14
Much more than an archive, a unique shared infrastructure
Much more than an archive, a unique shared infrastructure addressing the needs of *industry, research, culture and society as a whole*. 
Much more than an archive, a unique shared infrastructure addressing the needs of *industry, research, culture and society as a whole*. 

**Software Graph**

---

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

Software Heritage as a key infrastructure  13/12/2023  11 / 14
... a revolutionary infrastructure!

Much more than an archive, a unique shared infrastructure addressing the needs of industry, research, culture and society as a whole.
... a revolutionary infrastructure!

Much more than an archive, a unique shared infrastructure addressing the needs of industry, research, culture and society as a whole.

Software Graph

Software Blockchain

Open Science pillar

R. Di Cosmo  roberto@dicosmo.org (CC-BY 4.0)
... a revolutionary infrastructure!

Much more than an archive, a unique shared infrastructure addressing the needs of industry, research, culture and society as a whole.
Towards a resilient infrastructure

What about fire?

DALL-E's view of Library of Alexandria's fire

Thomas Jefferson, February 18, 1791

Welcoming ENEA's mirror

first institutional mirror

4 years of intense pathfinding work (huge thanks to David Douard, Stefano Ferriani and Simone/t_ta Pagnutti)

stepping stone to an European joint effort

R. Di Cosmo roberto@dicosmo.org (CC-BY 4.0)
Towards a resilient infrastructure

What about fire?

A lesson from the past

…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.

Thomas Jefferson, February 18, 1791
Towards a resilient infrastructure

What about fire?

A lesson from the past

...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.

Thomas Jefferson, February 18, 1791

Welcoming ENEA’s mirror

- **first** institutional mirror
- 4 years of intense **pathfinding** work (huge thanks to David Douard, Stefano Ferriani and Simonetta Pagnutti)
- stepping stone to
  - an European joint effort
Outline

1. Introduction
2. Software Heritage to the rescue
3. Library of Alexandria as an inspiration
4. The way forward
Bring together academia, industry, governments, communities

"to build a reference, global infrastructure for open and better software"
A rally flag for a grand vision

Bring together academia, industry, governments, communities

"to build a reference, global infrastructure for open and better software"

Software Heritage is the first brick …

- vendor neutral
- open source
- a worldwide initiative
- a long term initiative
A rally flag for a grand vision

Bring together academia, industry, governments, communities

"to build a reference, global infrastructure for open and better software"

Software Heritage is the first brick …
- vendor neutral
- open source
- a worldwide initiative
- a long term initiative

… that will enable
- archival, reference, integrity
- qualification, sharing and reuse
- a global software knowledge base
- responsible AI for code, and more
A rally flag for a grand vision

Bring together academia, industry, governments, communities
"to build a reference, global infrastructure for open and better software"

Software Heritage is the first brick …
- vendor neutral
- open source
- a worldwide initiative
- a long term initiative

… that will enable
- archival, reference, integrity
- qualification, sharing and reuse
- a global software knowledge base
- responsible AI for code, and more

We are making progress
Software Heritage is a key infrastructure in a long overdue undertaking

Today, we celebrate an important milestone on this path!
Learn more about Software Heritage
Learn more about Software Heritage

Annual report

5 years in 5 minutes

Evolution of our codebase

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