

Software Heritage: global archive to enable our digital future

Roberto Di Cosmo

Director, Software Heritage
Inria and Université de Paris Cité

13 December 2023



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

- 
- 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."

GPL Licence





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

GPL Licence

Hello World



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

GPL Licence

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
```



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

GPL Licence

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
```

Program (source code)

```
/* Hello World program */
#include<stdio.h>

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

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.”

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)

```
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     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 ...
              CADR   BURNBABY
```


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)

```
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      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 ...
              CADR    BURNBABY
```

Quake III source code (excerpt)

```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 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 * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
    // y = y * ( threehalfs - ( 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*)

```
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      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 ...
              CADR    BURNBABY
```

Quake III source code (*excerpt*)

```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 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 * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
    // y = y * ( threehalfs - ( 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.”

(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

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



(Open Source) 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



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

(Open Source) 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



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

(Open Source) 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



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](#)

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

Software source code is fragile

Endangered source code ...



A word cloud containing terms such as: damage, disaster, malicious, obsolete, attack, dependencies, aging, media, dangling, wear, corruption, encryption, format, deletion, reference, and storage.

- *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*

A global call for action

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 ...

A global call for action

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 Feb 2019

A global call for action

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

6 November 2018



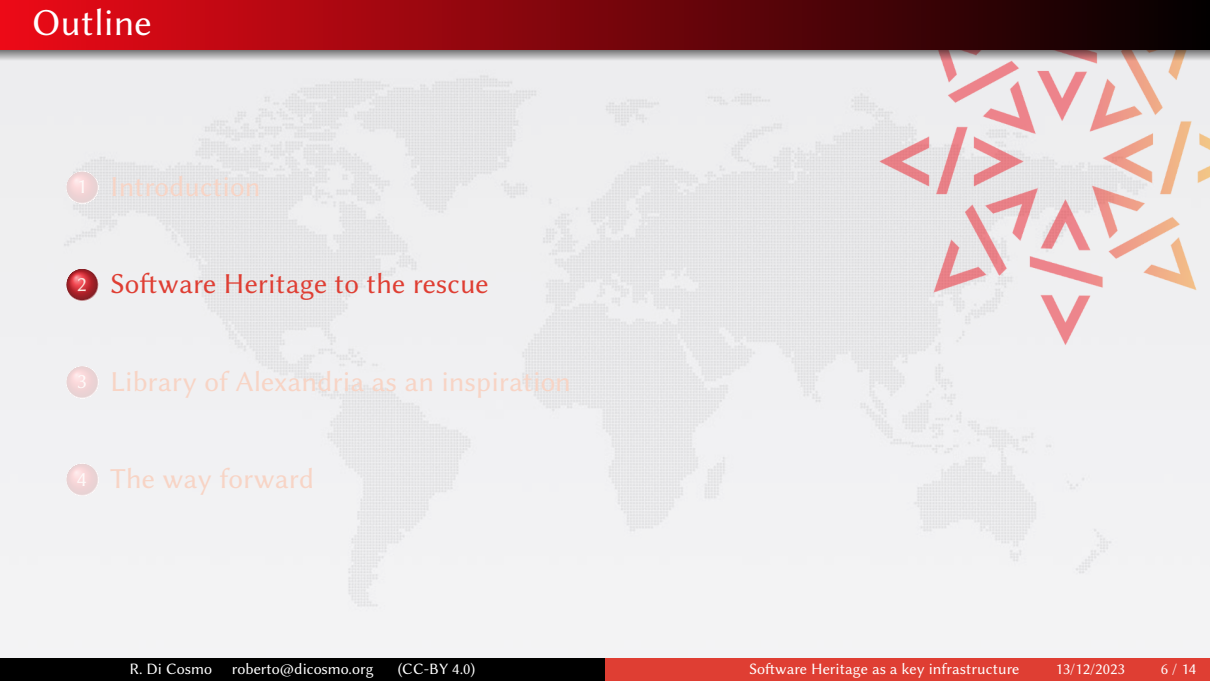
UNESCO, Inria, Software Heritage invite
40 international experts meet in Paris ...

“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;”

<https://en.unesco.org/foss/paris-call-software-source-code>



The call is published on Feb 2019

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



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

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



Software Heritage

THE GREAT LIBRARY OF 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



Software Heritage

THE GREAT LIBRARY OF 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



Software Heritage

THE GREAT LIBRARY OF 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

Sharing the vision



United Nations
Educational, Scientific and
Cultural Organization



And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors



Diamond sponsor



Platinum sponsors



Gold sponsors



Silver sponsors



Bronze sponsors



The largest software archive, a shared infrastructure

One infrastructure
open and shared



The largest software archive, a shared infrastructure

One infrastructure
open and shared

Cultural Heritage



Industry



Research



Public Administration



Software Heritage

The largest archive ever built



The largest software archive, a shared infrastructure

One infrastructure
open and shared



The largest archive ever built



Directories

13,486,743,792

Commits

3,555,307,942

Authors

63,835,635


Projects

262,077,475

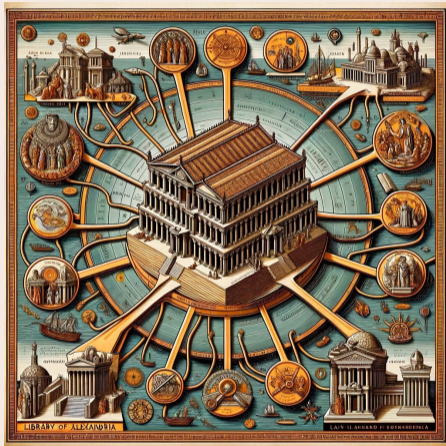
Releases

74,477,682

Bitbucket 2,012,133 origins	git 19,494 origins	R 21,486 origins
debian 129,217 origins	gn 6,424 origins	GitHub 152,282,093 origins
GitLab 3,989,638 origins	Guix 12,451 origins	GNU 354 origins
heptapod 1,096 origins	launchpad 356,873 origins	Maven 93,710 origins
NixOS 12,451 origins	npm 1,799,296 origins	Ubuntu 4,083 origins
Phabricator 185 origins	puthon 427,135 origins	SOURCEFORGE 308,970 origins

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

The Great Library of Alexandria



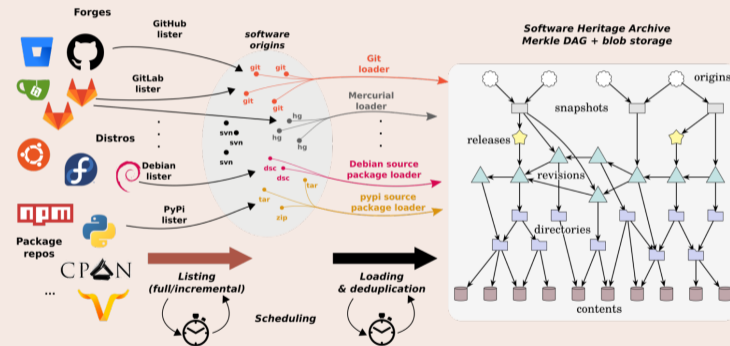
DALL-E view of Library of Alexandria

Built by the Ptolemys (3rd C BC):

- send scholars all over the world...
- ... to copy precious documents ...
- ... and bring them back under one roof!

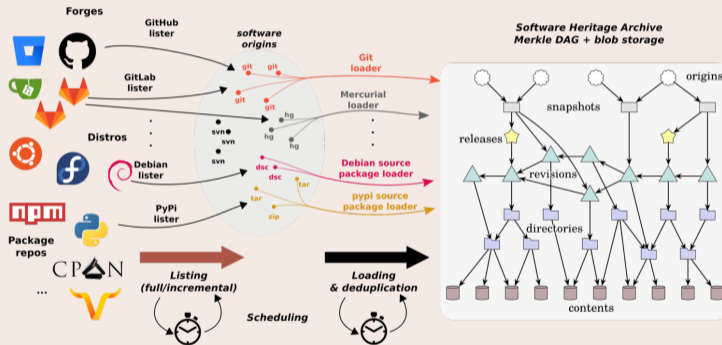
... to a modern Library of Alexandria for Software Source Code

Software Heritage under the hood



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

Software Heritage under the hood



Software Heritage by DALL-E



DALL-E's view

... a revolutionary infrastructure!

Much more than an archive, a unique shared infrastructure

... 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



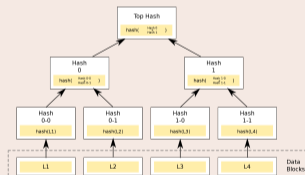
... a revolutionary infrastructure!

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



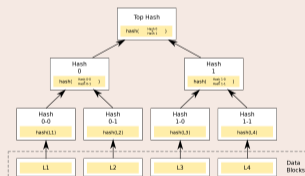
... a revolutionary infrastructure!

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

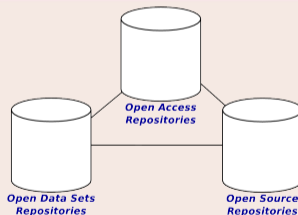
Software Graph



Software Blockchain



Open Science pillar



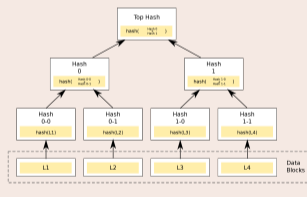
... a revolutionary infrastructure!

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

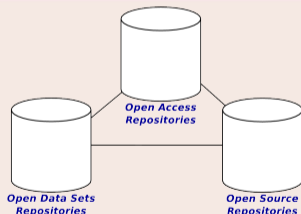
Software Graph



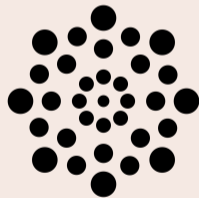
Software Blockchain



Open Science pillar



Big Code



What about fire?



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

What about fire?



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

What about fire?



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

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

A rally flag for a grand vision

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



 **Software Heritage**
THE GREAT LIBRARY OF SOURCE CODE

Collecting, preserving
and sharing software
source code since 2013



5 years in 5 minutes

[Link](#)



Learn more about Software Heritage

Annual report



Collecting, preserving
and sharing software
source code since 2013

5 years in 5 minutes

[Link](#)



Evolution of our codebase

[Link](#)

