Software Heritage

preserving our software commons

Roberto Di Cosmo
Inria and Université de Paris

18 June 2021
RESAW 2021 Keynote

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Introduction
2. Software Source Code is Precious Knowledge
3. An Endangered Knowledge
4. Meet Software Heritage
5. Demo time!
6. Preserving the history of our software commons
7. The road ahead

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- *30 years* of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- *20 years* of Free and Open Source Software
- *10 years* building and directing structures for the common good

1999 *DemoLinux* – first live GNU/Linux distro

2007 *Free Software Thematic Group*
150 members 40 projects 200Me

2008 *Mancoosi project* www.mancoosi.org

2010 *IRILL* www.irill.org

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

Outline

1. Introduction
2. Software Source Code is Precious Knowledge
3. An Endangered Knowledge
4. Meet Software Heritage
5. Demo time!
6. Preserving the history of our software commons
7. The road ahead

What is *Source Code*?

"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* matters

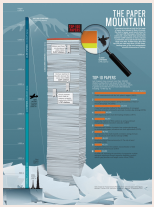
Key mediator for accessing *all* information (c) Banski

Information is a main pillar of our modern societies.

Absent an ability to correctly interpret digital information, we are left with [...] "rotting bits" [...] of no value.

Vinton G. Cerf IEEE 2011

Software is *an essential component* of modern scientific research



[...] the vast majority describe experimental methods or software that have become essential in their fields.

Top 100 papers (Nature, October 2014)

Sometimes, if you dont have the software, you dont have the data

Christine Borgman, Paris, 2018

Sofware embodies our *Knowledge* and *Cultural Heritage*

Software *Source Code* is part of our cultural heritage

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     CODE500         # ASTRONAUT:    PLEASE CRANK THE
                TC      BANKCALL        #               SILLY THING AROUND
                CADR    GOPERF1
                TCF     GOTOPOOH        # 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 ...
                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

"Source code provides a view into the mind of the designer."

Source Code and the commons

Definition (Commons)

The commons is the cultural and natural resources accessible to all members of a society, including natural materials such as air, water, and a habitable earth. These resources are held in common, not owned privately. https://en.wikipedia.org/wiki/Commons

Definition (Software Commons)

The software commons consists of all computer software which is available at little or no cost and which can be altered and reused with few restrictions. Thus *all open source software and all free software are part of the [software] commons.* [...]

https://en.wikipedia.org/wiki/Software_Commons

Outline

1. Introduction
2. Software Source Code is Precious Knowledge
3. An Endangered Knowledge
4. Meet Software Heritage
5. Demo time!
6. Preserving the history of our software commons
7. The road ahead

Spread all around

Fashion victims

- many disparate development platforms
- a myriad places where distribution may happen
- projects tend to migrate from one place to another over time

One place…

… where can we find, track and search *all* source code?

Fragile

Like all digital information, FOSS is fragile

- inconsiderate and/or malicious code loss (e.g., Code Spaces)
- business-driven code loss (e.g., Gitorious, Google Code, Bitbucket)
- for obsolete code: physical media decay (data rot)

If a website disappears you go to the Internet Archive…

where do you go if (a repository on) GitHub or GitLab goes away?
and what about all the landmark legacy source code that is rotting away?

We are at a turning point

Looking at the past

- a lot of old software misplaced, lost, or behind barriers, but…
- most founding fathers are still here, and willing to share
- urgent to collect their knowledge

Only a few years left.

Looking at the future

- software development and use skyrockets: more programmers, and more code!
- essential to provide a universal platform for all the future software source code

Every year that goes by makes the problem worse.

it is urgent to take action!

Outline

1. Introduction
2. Software Source Code is Precious Knowledge
3. An Endangered Knowledge
4. Meet Software Heritage
5. Demo time!
6. Preserving the history of our software commons
7. The road ahead

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 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

Platinum sponsors

Gold sponsors

Silver sponsors

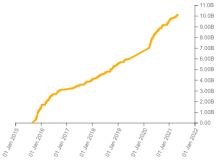
Bronze sponsors

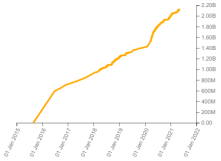
The largest software archive, a shared infrastructure

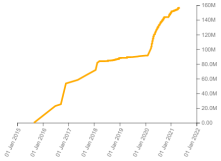
Cultural Heritage | **Industry** | **Research** | **Education**

Software Heritage

Source files	Commits	Projects
10,162,254,604	2,128,899,595	156,417,833







Directories	Authors	Releases
8,477,359,190	42,598,760	17,293,952

Growing adoption in Academia (selection)

HAL software curated deposit workflow

Curated Archiving of Research Software Artifacts
International Journal of Digical Curation, 2020

Reference archive for swmath.org

swMATH
an information service for mathematical software

See *code* links, e.g. SemiPar package

IPOL (image processing)

- archive (deposit)
- reference
- BibLaTeX

eLife (life sciences)

- archive (save code now)
- reference

JTCAM (Mechanics)

- instructions for authors
- biblatex-software in journal LaTeX class

Policy: France

National Plan Open Science

Policy: Europe

EOSC SIRS report

- SWHIDs
- archive

Guidelines

- summary
- ICMS 2020

Software Heritage Identifiers (SWHID)

link to full docs

An emerging standard

- in Linux Foundation's SPDX 2.2
- IANA registered, WikiData property P6138

Examples:

- Apollo 11 AGC excerpt,
- Quake III rsqrt

Outline

1. Introduction
2. Software Source Code is Precious Knowledge
3. An Endangered Knowledge
4. Meet Software Heritage
5. Demo time!
6. Preserving the history of our software commons
7. The road ahead

Focus on Archive and Reference

- The Apollo 11 AGC source code example
- Browse the archive
- Trigger archival of your preferred software in a breeze
- Get and use SWHIDs (full specification available online)
- Example use in a research article: compare Fig. 1 and conclusions
  - in the 2012 version
  - in the updated version using SWHIDs and Software Heritage
- Cite software with the biblatex-software style from CTAN
- Example in a real journal: an article from IPOL

Recent news, and a lesson to be learned

Saving 250.000 endangered repositories…

- summer 2019: BitBucket announce Mercurial VCS phase out
- fall 2019: Software Heritage teams up with Octobus (funded by NLNet, thanks!)
- july 2020: BitBucket erases *250.000* repositories
- august 2020: bitbucket-archive.softwareheritage.org is live

… preserving the web of knowledge (Tweet is here)

Bottomline

explicit deposit is important, …

… and we must promote it…

… but will never be enough.

(think also of all software dependencies!)

Summing up: a revolutionary infrastructure *designed for source code*

Software Heritage

www.softwareheritage.org

global source code _archive_ — *Library of Alexandria of source code*

- harvest *all* software source code
- *on demand harvesting* and *curated deposit*

universal intrinsic identifiers

SWHID standard is independent of version control systems

uniform data model, _full graph_ of development history

enables large scale, big code research

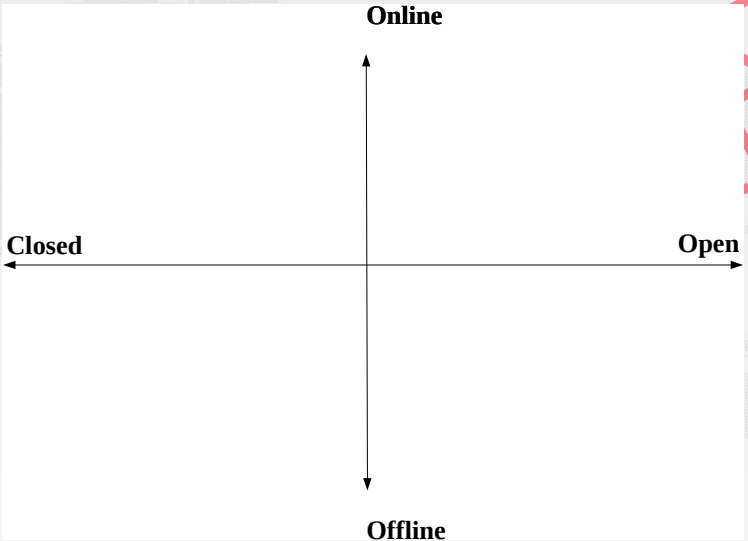
infrastructure for Open Science

base layer for software source code in the *Open Science architecture*

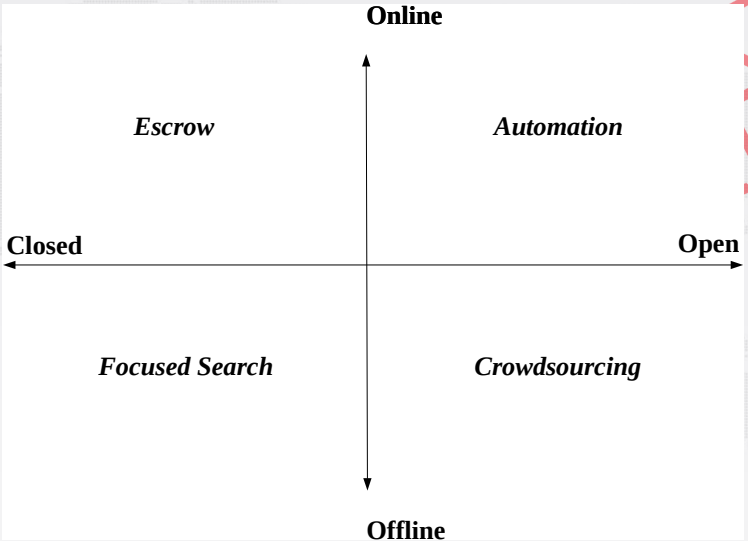
Outline

1. Introduction
2. Software Source Code is Precious Knowledge
3. An Endangered Knowledge
4. Meet Software Heritage
5. Demo time!
6. Preserving the history of our software commons
7. The road ahead

All the source code!

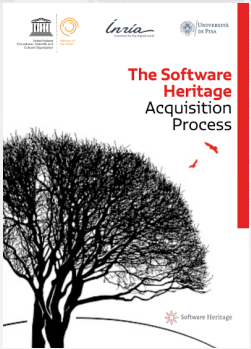


All the source code: strategy



Paris Call on Software Source Code

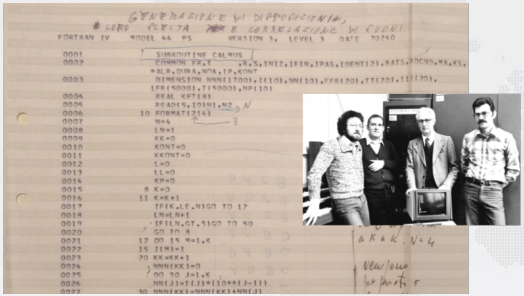
"[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive"



- Rescue Legacy Software from different media
- Curate the code
  - reconstruct the development history
    - *Software Heritage - GitHub* work on fixing git
  - collect the metadata
- Archive in Software Heritage

UNESCO, UniPi and Software Heritage collaboration worldwide scope

An example: TAUmus, from Pisa (70's)



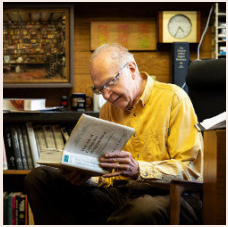
- The control code for the music synthesizer TAU2
- FORTRAN II and TAUmus command language
- Istituto di Elaborazione dell'Informazione CNR
- Group led by the late M° P. Grossi
  - Le Sacre du Printemps (ABSTRACT)

Outline

1. Introduction
2. Software Source Code is Precious Knowledge
3. An Endangered Knowledge
4. Meet Software Heritage
5. Demo time!
6. Preserving the history of our software commons
7. The road ahead

Calling for more History of Computer Science

Communications of the ACM, February 2021



"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."

Let's Not Dumb Down the History of Computer Science
Donald E. Knuth, Len Shustek
https://doi.org/10.1145/3442377

A unique opportunity

- most of the creators are still here: we can talk to them!
- Software Heritage provides a key infrastructure for software historians

Stay tuned

great news for software stories in a few months...

You may help!

Foster adoption and best practices

- archive and reference relevant source code (save code now, and deposit),
- rescue and preserve landmark legacy source code with SWHAP,
- use Software Heritage in articles, journals, and books

Engage with Software Heritage as an individual

- the ambassador program is open:
  - learn directly from the core team, share with your community

Engage with Software Heritage as an organization

- join as a member/sponsor
- contribute to the preservation mission

Questions?

References

R. Di Cosmo, *A revolutionary infrastructure for Open Source*, 2021, EU Software Forum (slides) (video)

EOSC SIRS Task Force, *Scholarly Infrastructures for Research Software*
2020, European Commission, (10.2777/28598)

R. Di Cosmo, *Archiving and Referencing Source Code with Software Heritage*
ICMS 2020 (10.1007/978-3-030-52200-1_36)

R. Di Cosmo, M. Gruenpeter, S. Zacchiroli
Referencing Source Code Artifacts: a Separate Concern in Software Citation,
CiSE 2020 (10.1109/MCSE.2019.2963148) (hal-02446202)

P. Alliez, R. Di Cosmo, B. Guedj, A. Girault, M.-S. Hacid, A. Legrand and N. Rougier
Attributing and referencing (research) software: Best practices and outlook from Inria,
CiSE 2020 (10.1109/MCSE.2019.2949413) (hal-02135891)

J.F. Abramatic, R. Di Cosmo, S. Zacchiroli, *Building the Universal Archive of Source Code*,
CACM, October 2018 (10.1145/3183558)