Preserving the world's software history

why we must, and how we can

Roberto Di Cosmo

Director, Software Heritage
Inria and Université de Paris

13/06/2022
Ljubjana 2022

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

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
2021	*EOSC Task Force on Infrastructures for Software*, European Union

Outline

Software *Source Code* is Precious 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     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)

```c
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."

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

6 November 2018

PARIS CALL
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

"[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"
https://en.unesco.org/foss/paris-call-software-source-code

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!

but the clock is ticking…

Outline

damage
disaster
reference
storage
deletion
media malicious
aging attack obsolete
tear dependencies
dangling
wear
corruption
encryption
format

Like all digital information, FOSS is fragile

- link rot: projects are created, moved around, removed
- business-driven code loss (e.g., Gitorious, Google Code, Bitbucket)
- data rot: physical media with legacy software decay

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 early creators 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!

A complex mission, a noble endeavour

Many challenges lie in front of us

Technical

- recover software (source code)
- archive and reference it
- make it run again

Legal

- licence
  - get the licence rights…
  - get the right license

Methodological

- tell the story of the people
  - get the relevant documents
  - get the oral history

Financial and Organizational

- funding model for archives/museums
- build an international network
- share resources and knowledge

let's take action!

Outline

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

| Cultural Heritage | Industry | Research | Public Administration |

Software Heritage

Source files	Commits	Projects
12,032,627,304	2,536,918,821	173,242,749

Directories	Authors	Releases
9,946,192,395	47,334,620	31,763,605

Technology

- transparency and FOSS
- replicas all the way down

Content (billions!)

- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- multi-stakeholder

A walkthrough

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

- Example in a journal: an article from IPOL
- Curated deposit in SWH via HAL, see for example: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, …
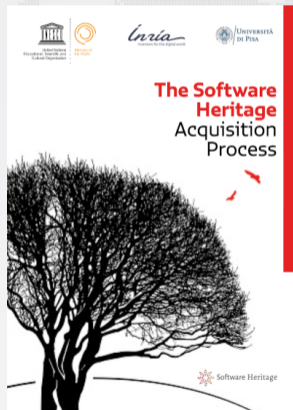- Rescue landmark legacy software, see the SWHAP process with UNESCO

Outline

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"

The Software Heritage Acquisition Process

- **Rescue** Legacy Software from different media
  - physical
  - digital
    - legacy / unsupported
    - recent / supported

- **Curate** the code
  - reconstructing the development history
  - collecting metadata

- And **illustrate** with dedicated presentations

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"

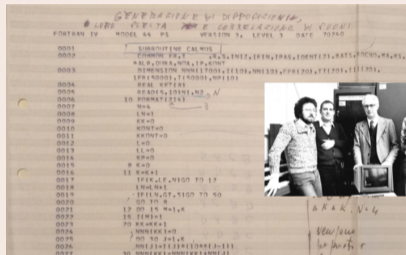SOFTWARE STORIES
Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

The Pisa Collection

Stories of landmark legacy code (Beta Version)

3 STORIES IN THIS COLLECTION

Search Collection...

TAUmus

- Expand the SWHAP scope to
  - documents
  - media (videos, pictures, images, etc.)
  - oral history

- Preserve and Present all this material

- Share process and tools (all open source!)
  - with museums, archives and all interested parties

see this live on the Software Stories website, and get *the guide*

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

Electronic music in Pisa: group led by the late M° P. Grossi

- Control code of the music synthesizer TAU2
- FORTRAN II, TAUmus command language
- Istituto di Elaborazione dell'Informazione CNR
- e.g. Le Sacre du Printemps (ABSTRACT)

See this live

- the archived SWHAP repository
- and its Software Story

Outline

Sharing the vision

And many more ...
www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Diamond sponsor

Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors

Foster adoption and best practices

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

Engage with Software Heritage

- join the ambassador program, like Borut and Bostjan
- become a member/sponsor

Contribute to the preservation mission

- support museums and archives
- create an international network

Questions?