Preserving our Software Heritage and its Stories: why and how

Roberto Di Cosmo
Bologna

Director, Software Heritage
Inria and Université de Paris Cité

December 1st 2022

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

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)

```
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."

Calling for preservation: UNESCO

UNESCO, Inria, Software Heritage invite

40 international experts meet in Paris …

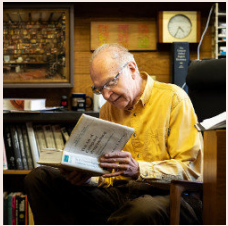
The call is published on Feb 2019

"[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

Calling for preservation: Donald Knuth and Len Shustek

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

Some popular approaches

A - Since the ~~1970's~~ 1990's

.zip or .tar file on:

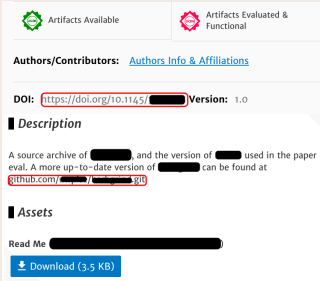
- ~~ftp server~~
- web page
- document archive (+ DOI)

B - Since the 2000's

Rely on *software forges*

- institutional or project ones
- free commercial ones: BitBucket, GitHub, GitLab, …

C: a mix of the two



Can get no satisfaction…

A *Poor user experience*

B *Preservation?*

C Can do better

Forges are *not* archives!

2015: the first big bad news

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

- broken links in the web of knowledge (my papers too)

Big bad news keep coming in

- summer 2019: BitBucket announces Mercurial VCS sunset
- july 2020: BitBucket erases *250.000+* repositories (including research software)
- summer 2022: GitLab.com considers erasing all projects that are inactive for a year

In Academia too!

- 2021: Inria's old gforge is unplugged… breaks the Opam build chain for OCaml

We need a universal archive of software source code: now we have one!

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 and share all software source code

Research infrastructure

enable analysis of all software source code

Universal software archive, principled http://bit.ly/swhpaper

One infrastructure
open and shared

Cultural Heritage | Industry | Research | Public Administration

Software Heritage

Largest archive

Technology

- transparency and FOSS
- replicas all the way down

Content (billions!)

- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- multi-stakeholder

An international, non profit initiative built for the long term

Sharing the vision

United Nations
Educational, Scientific and
Cultural Organization

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

Donors, members, sponsors

Diamond sponsor

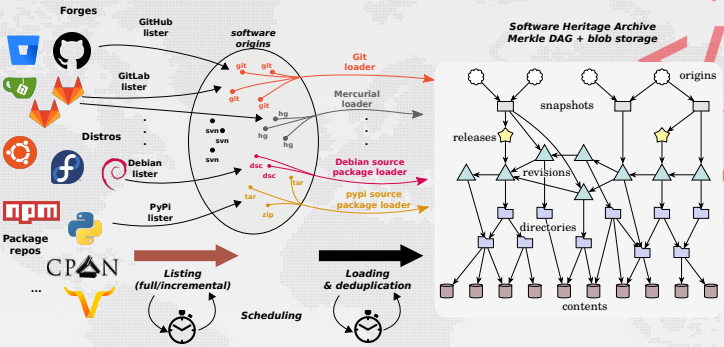
Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors

A peek under the hood: a universal archive



Global development history permanently archived in a uniform data model

- over 12 billion unique source files from over 180 million software projects
- ~1PB (uncompressed) blobs, ~25 B nodes, ~350 B edges

Intrinsic Identifiers for software artefacts

Software Heritage Identifiers (SWHID)

link to full docs

25+B intrinsic, decentralised, cryptographically strong identifiers, SWHIDs

Emerging standard : Linux Foundation SPDX 2.2; IANA registered; WikiData P6138

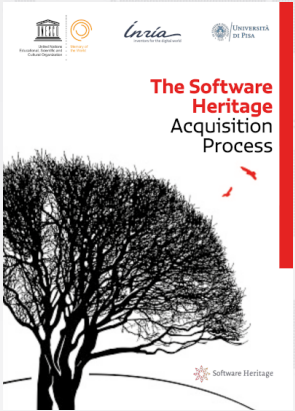
Full fledged *source code references* for reproducibility

Examples: Apollo 11 AGC excerpt, Quake III rsqrt; Guidelines available, see ICMS 2020

Outline

Paris Call on Software Source Code

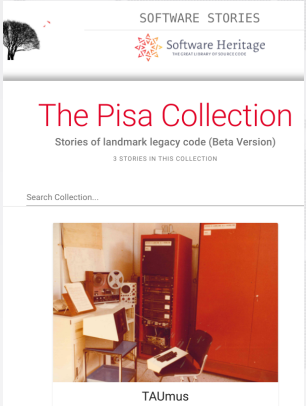
"[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive"



- 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"

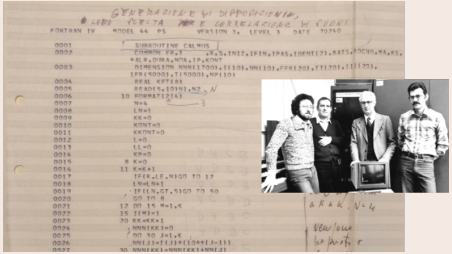


- Expand the SWHAP scope to
  - documents
  - media (videos, pictures, images, etc.)
  - oral history
- Preserve and Present all this material
- Share process and tools (all open source!)
  - with museums, archives and all interested parties

see this live on the Software Stories website, and get *the guide* and the SWHAP Days hybrid event proceedings, 19 and 20 october 2022

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

Meet the team

Laura Bussi
PhD candidate
in Computer Science
SWHAP
University of Pisa

Carlo Montangero
Computer Scientist, SWHAP
University of Pisa

Kenneth Seals-Nutt
Computer Scientist
& Software Engineer
Co-founder of ScienceStories.io

Roberto Di Cosmo
Computer Scientist
Founder of Software Heritage

Elisabetta Mori
Historian of Computing
SWH Visiting Scientist

Katherine Thornton
Information Scientist
Co-founder of ScienceStories.io

Morane Gruenpeter
Software engineer
Project Manager

Guido Scatena
Computer Scientist, SWHAP
ISPRA

A proposal for a working agenda

Search and find software source code associated to *landmark research articles*

Reconstruct development history, *archive in SWH*

Link publications to the source code using the SWHID identifier

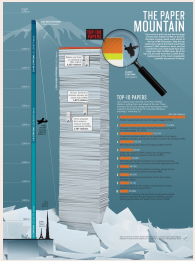
Collect oral and documentary history around it, and build a Software Story

Connect with all the relevant history collections

Outline

Source code history for Open Science

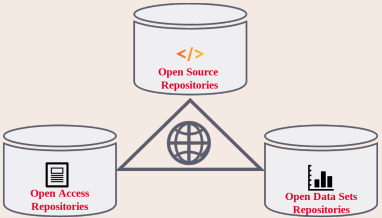
Software powers modern research



[…] software […] essential in their fields.
Top 100 papers (Nature, 2014)

Sometimes, if you dont have the software, you dont have the data
Christine Borgman, Paris, 2018

A key pillar: software (source code)



The links in the picture are important

Nota Bene

software may be a *tool*, a *research outcome* and a *research object*

access to the *source code* is essential!

Preserving the history of source code is important for *reproducibility*

Using Software Heritage for Open Science

- Browse (e.g. Apollo 11, and your work may be already there !)
- Trigger archival, use the updateswh browser extension (GitHub action available too)
- Get and use SWHIDs (full specification available online)
- Cite software with biblatex-software package from CTAN
  - Overleaf ACMART template available
- Example in journals: article from IPOL
- Example with Parmap: devel on Github, archive in SWH, curated deposit in HAL
- Extracting all the software products for Inria, for CNRS, for CNES, for LIRMM or for Rémi Gribonval using HalTools
- Curated deposit in SWH via HAL, see for example: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, ...
- Example use in research articles:
  - compare Fig. 1 and conclusions in the 2012 version and the updated version
  - SWHID in a replication experiment

Let's do it right from the start

Archiving and Referencing

For all source code used in research (*yes, even small scripts!*)

- ensure it is archived in Software Heritage (see save code now)
- get the proper SWHID for your software (see detailed HOWTO)
- add it to research articles for reproducibility (see detailed HOWTO)

Describing and Citing/Crediting

For software projects, go the extra mile:

- add proper metadata (e.g. codemeta.json, see the codemeta generator)
- cite software (e.g. using biblatex-software, in CTAN, TeXLive and acmart)
- index on par with publications (see the french portal HAL)

ACM action item

connect ACM DL and Badging program with Software Heritage

Focus on Academia: growing adoption (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 FOR OPEN SCIENCE

National Plan for Open Science and Research Infrastructures

Policy: Europe

EOSC SIRS report

- SWHIDs
- archive

Guidelines

Software Heritage
1 Prepare your public repository README, AUTHORS & LICENSE files
2 Save your code http://save.softwareheritage.org/
3 Reference your work (full repository, specific version or code fragment)

- summary
- ICMS 2020

Preserving for the long term

Thomas Jefferson, February 18, 1791

…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.

First institutional mirror will be in Italy

Italian National Agency for New Technologies, Energy and Sustainable Economic Development

- agreement in 2019
- deployment ongoing
- stepping stone to an European joint effort

Outline

Join a growing, active community

The first five years in just five minutes

Ambassadors, news, blog, media

- meet the ambassadors
- subscribe to the newsletter
- read the blog
- follow @swheritage

Questions? 5 minutes recap https://youtu.be/Ez4xKTKJO2o

A long way to go: it is urgent to get started!

References (see https://www.softwareheritage.org/publications)

- Morane Gruenpeter, Roberto Di Cosmo, Katherine Thornton, Kenneth Seals-Nutt, Carlo Montangero, Guido Scatena
  Software Stories for landmark legacy code, Inria TR, 2022
  https://hal.archives-ouvertes.fr/hal-03483982
- 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)
- Laura Bussi, Roberto Di Cosmo, Carlo Montangero, Guido Scatena
  The software heritage acquisition process
  UNESCO, Università di Pisa, Inria, 2019
- J.F. Abramatic, R. Di Cosmo, S. Zacchiroli, *Building the Universal Archive of Source Code*,
  CACM, October 2018 (10.1145/3183558)