Software Source Code in Research and Open Science

Roberto Di Cosmo
Director, Software Heritage

July 2020
1. Software Source Code is knowledge
2. Software Heritage
3. Demo time!
4. The way forward
Software source code: *human readable and executable knowledge*

Harold Abelson, Structure and Interpretation of Computer Programs (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 BANCKALL  # SILLY THING AROUND
           CADR GOPERF1
           TCF GOTOP00H  # TERMINATE
           TCF P63SPOT3  # PROCEED  SEE IF HE'S LYING

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

Len Shustek, Computer History Museum (2006)

“Source code provides a view into the mind of the designer.”

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 = 0x5f3759dfL * ( 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;
}
```
Source code is special (software is not data)

Software evolves over time

- projects may last decades
- the development history is key to its understanding

Complexity

- millions of lines of code
- large web of dependencies
  - easy to break, difficult to maintain
- sophisticated developer communities

A vast part is not research software

- industry and communities drive standards, build the necessary support layers
Versioning, granularity

Project  “Inria created OCaml and Scikit-learn”
Release  “2D Voronoi Diagrams were introduced in CGAL 3.1.0”
Precise state of a project  “This result was produced using commit 0064fbd…”
Code fragment  “The core algorithm is in lines 101 to 143 of the file parmap.ml contained in the precise state of the project corresponding to commit 0064fbd…”

Authors can have multiple roles:
- Architecture, Management, Development, Documentation, Testing, …
Software Source code: pillar of Open Science

Three pillars of Open Science

- Open Data Sets Repositories
- Open Access Repositories
- Open Source Repositories

A plurality of needs

**Researcher**
- archive and reference software used in articles
- find useful software
- get credit for developed software
- verify/reproduce/improve results

**Laboratory/team**
- track software contributions
- produce reports / web page

**Research Organization**
- know its software assets
- technology transfer
- impact metrics
<table>
<thead>
<tr>
<th>What is at stake</th>
<th>in increasing order of difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archival</td>
<td>Research software artifacts must be properly archived make sure we can retrieve them (<em>reproducibility</em>)</td>
</tr>
<tr>
<td>Identification</td>
<td>Research software artifacts must be properly referenced make sure we can identify them (<em>reproducibility</em>)</td>
</tr>
<tr>
<td>Metadata</td>
<td>Research software artifacts must be properly described make it easy to discover them (<em>visibility</em>)</td>
</tr>
<tr>
<td>Citation</td>
<td>Research software artifacts must be properly cited (<em>not the same as referenced</em>)! to give credit to authors (<em>evaluation!</em></td>
</tr>
</tbody>
</table>

We need infrastructures *designed for* software source code: now we have one!
Outline

1. Software Source Code is knowledge
2. Software Heritage
3. Demo time!
4. The way forward
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

Software Heritage, in a nutshell
Addressing the four needs

Archive (8B+ files, 130M+ projects!)

- save.softwareheritage.org
- deposit.softwareheritage.org

Reference (20 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs

Now supported in SPDX 2.2, Wikidata etc.

Describe

- Intrinsic metadata from source code
- Contributed the Codemeta generator

Cite

- Contributed software citation style biblatex-software, v 1.2-2 now on CTAN
Outline

1. Software Source Code is knowledge
2. Software Heritage
3. Demo time!
4. The way forward
A walkthrough

- Browse the archive
- Get and use SWHIDs (full specification available online)
- 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 use in a research article: extensive use of SWHIDs in a replication experiment
- Trigger archival of your preferred software in a breeze
- 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

1. Software Source Code is knowledge
2. Software Heritage
3. Demo time!
4. The way forward
An international, non profit initiative built for the long term

Sharing the vision

United Nations Educational, Scientific and Cultural Organization

Donors, members, sponsors

Platinum sponsors

Microsoft

Intel

Société Générale

Huawei

Gold sponsor

openinventionnetwork

Université de Paris

Silver sponsors

Creative Commons

Mozilla

GitHub

Google

Università di Pisa

Bronze sponsors

Open Source Initiative

FLOSSID

Bell Labs

UQÀM

Université de Québec à Montréal

Archiving and Referencing source code CC-BY 4.0 July 2020
Come in, we’re open!

**Software Heritage**
- *universal* source code archive
- *intrinsic* identifiers (SWHIDS)
- *open, non profit,* long term
- *infrastructure* for Open Science

**You can help improve science!**
- *use* SWH and *save* relevant source code
- *build on* SWH (see swmath.org and ipol.im)
- *contribute* to SWH: *it is open source*
- spread the word

---

Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchiroli
*Building the Universal Archive of Source Code*, CACM, October 2018 (*10.1145/3183558*)

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

Pierre Alliez, Roberto Di Cosmo, Benjamin Guedj, Alain Girault, Mohand-Said Hacid, Arnaud Legrand and Nicolas Rougier
*Attributing and referencing (research) software: Best practices and outlook from Inria*, CiSE 2020 (*10.1109/MCSE.2019.2949413*) (hal-02135891)
“The source code for a work means the preferred form of the work for
making modifications to it.”

Hello World

Program (source code)
/* Hello World program */
#include<stdio.h>
void main()
{
    printf("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
Software Heritage

- Source files: 7,309,842,258
- Commits: 1,590,436,149
- Projects: 107,875,943
A revolutionary infrastructure for software source code

The **graph** of Software Development

All software development with its history, in a single graph ...

The **blockchain** of Software Development

... a single Merkle graph, with *intrinsc ids* for traceability

A **pillar** of Open Science

Reference **archive** of Research Software

Reference platform for **Big Code**

One uniform data structure enables *massive* machine learning for quality, cybersecurity, etc.
Outline

5 Milestones

6 Breaking news

7 A bit of tech info

8 The way forward
Some key dates

Summer 2015
The collection starts: first server, (very) early prototype

June 30th 2016
Public unveiling, with the first sponsors: Microsoft and DANS

April 3rd 2017
Unesco - Inria agreement on software access and preservation.

June 7th 2018
Opening the archive to the world

December 7th 2018
Starting the mirror network

February 26th 2019
Publication of the expert meeting Paris Call on Software Source Code
UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …

It’s an important *policy tool*, already referenced and used … yes, you can sign it!

Archiving public code

Public sector source codes
This website lets you browse some of the source codes opened by public bodies. If your source code is not referenced on this website, send us a link to your repository.

Source code repositories  Organizations or groups  Figures  Free search

License  Language  Forks only  Hide archives  With a description  With known license  3669 repositories

<table>
<thead>
<tr>
<th>Repository / group</th>
<th>Archive</th>
<th>Description</th>
<th>Updated</th>
<th>Forks</th>
<th>Stars</th>
<th>Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>medle / SocialGouv</td>
<td>✨</td>
<td>MedLé : plateforme permettant aux établissements de santé de déclarer leur activité médico-légale</td>
<td>11/8/2019</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>reseauxchaleur / dreall-datalab</td>
<td>✨</td>
<td></td>
<td>11/8/2019</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

https://code.etalab.gouv.fr
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.

Welcoming ENEA

- first institutional mirror
- increased resilience
- AI infrastructure for researchers
- stepping stone to an European joint effort
The Software Heritage Acquisition Process (SWHAP)

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”

SWHAP : an important step forward

- detailed guidelines to curate landmark legacy source code and archive it on Software Heritage
- intense cooperation with Università di Pisa and UNESCO
- open to all, we’ll promote it worldwide

https://www.softwareheritage.org/swhap
Outline

5 Milestones

6 Breaking news

7 A bit of tech info

8 The way forward
Global development history permanently archived in a unique git-like Merkle DAG

- ~400 TB (uncompressed) blobs, ~20 B nodes, ~280 B edges
Sharing the vision

An international, non profit initiative built for the long term

Donors, members, sponsors

www.softwareheritage.org/support/testimonials

@rdicosmo Roberto Di Cosmo @swheritage

Archiving and Referencing source code CC-BY 4.0 July 2020 10 / 12
CNRS and Software Heritage Working together

Resources

- as a top SWH sponsor, contribute to the roadmap
  - funding (see https://sponsors.softwareheritage.org)
  - compute and storage infrastructure
  - engineers and researchers

Collaboration

- training and adoption
  - across disciplines
  - at the international level

- strengthen SWH interoperability with HAL

Strategy

- work together to establish an international infrastructure
Learn more

Roberto Di Cosmo. *Archiving and referencing source code with Software Heritage*
ICMS 2020, preprint: hal-02526083

Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchiroli
*Building the Universal Archive of Source Code*, Communications of the ACM, October 2018

P. Alliez, R. Di Cosmo, B. Guedj, A. Girault, M. Hacid, A. Legrand, N. Rougier

Roberto Di Cosmo, Morane Gruenpeter, Stefano Zacchiroli
*Referencing Source Code Artifacts: a Separate Concern in Software Citation*, Computing in Science & Engineering, 2020, ISSN: 1521-9615

Roberto Di Cosmo, Marco Danelutto