Archiving and Referencing all the software source code

Roberto Di Cosmo
Director, Software Heritage

ICMS 2020
1. Software Source Code is knowledge
2. Software Heritage
3. Demo time!
4. Conclusion
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)

```assembly
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 GOTOPO0M # 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 threehalves = 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 * ( threehalves - ( x2 * y + y ) ); // 1st iteration
  // y = y * ( threehalves - ( 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.”
Three pillars of Open Science

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>Category</th>
<th>Requirement</th>
<th>Make Sure We Can</th>
<th>(Reproducibility)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archival</td>
<td>Research software artifacts must be properly</td>
<td>archived</td>
<td>retrieve them</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(reproducibility)</td>
</tr>
<tr>
<td>Identification</td>
<td>Research software artifacts must be properly</td>
<td>referenced</td>
<td>identify them</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(reproducibility)</td>
</tr>
<tr>
<td>Metadata</td>
<td>Research software artifacts must be properly</td>
<td>described</td>
<td>discover them</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(visibility)</td>
</tr>
<tr>
<td>Citation</td>
<td>Research software artifacts must be properly</td>
<td>cited</td>
<td>give credit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>to authors</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(evaluation)</td>
</tr>
</tbody>
</table>

We need an infrastructure **designed for** software source code now we have it!
Outline

1. Software Source Code is knowledge
2. Software Heritage
3. Demo time!
4. Conclusion
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
full development history permanently archived!
over 8 billions unique source files from 130+ million origins
Software Heritage Identifiers (SWHID)

An emerging standard
- in Linux Foundation’s SPDX 2.2
- IANA registered, WikiData property P6138

Examples:
- Apollo 11 AGC excerpt,
- Quake III rsqrt
## Archive
- a *universal* archive: collects *all* software, not only academic software
- *harvests* source code worldwide *(8B+ files from 130M+ projects in July 2020)*
- your software may be there already… if not, please *save its code now!*

## Reference
- **SWHID:** *intrinsic, decentralised, cryptographically strong* identifiers
- enhance articles with *source code references* for reproducibility

## Cite
- **biblatex-software:** a dedicated bibliographic style for software!

**Detailed guidelines in the paper and online!**
Outline

1. Software Source Code is knowledge
2. Software Heritage
3. Demo time!
4. Conclusion
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. Conclusion
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

Platinum sponsors

Microsoft  Intel  SOCIETE GENERALE  HUAWEI

Gold sponsor

openinventionnetwork

Silver sponsors

Université de Paris

Bronze sponsors

DANS  FOSSID  NOOK  Bell Labs

Archiving and Referencing source code  CC-BY 4.0  July 2020 0 / 11
The way forward

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

**You can help improve science!**
- *use* SWH (see swmath.org and ipol.im)
- *save* relevant source code
- *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)