Software Heritage: a common infrastructure to preserve our Software Commons

Roberto Di Cosmo
Inria and Université de Paris

18/09/2020, EsLibre
Outline

1. **Software Source Code: precious part of our Knowledge**

2. **Meet Software Heritage**

3. **Demo time!**

4. **Building for the long term ...**

5. **... and firefighting too!**
“Programs must be written for people to read, and only incidentally for machines to execute.”

Apollo 11 source code (excerpt)

```
P63SP0T3
  CA  BIT6 # IS THE LR ANTENNA IN POSITION 1 YET
  EXTEND
  RAND  CHAN33
  EXTEND
  BZF  P63SP0T4 # BRANCH IF ANTENNA ALREADY IN POSITION 1
  CAF  CODE500 # ASTRONAUT: PLEASE CRANK THE
  TC  BANKCALL # SILLY THING AROUND
  CADR  GOPERF1
  TCF  GOTOPOOM # TERMINATE
  TCF  P63SP0T3 # PROCEED SEE IF HE’S LYING
P63SP0T4
  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 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

“Source code provides a view into the mind of the designer.”
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](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](https://en.wikipedia.org/wiki/Software_Commons)

Source code: part of our commons … pillar of Open Science! *(would require another talk)*

Precious, endangered **executable** and **human readable** knowledge

key people **passing away**, platforms (GoogleCode, Gitorious, etc.) closing down …
1. Software Source Code: precious part of our Knowledge

2. Meet Software Heritage

3. Demo time!

4. Building for the long term …

5. … and firefighting too!
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
Largest software archive, principled


Technology
- transparency and FOSS
- replicas all the way down

Content (billions!)
- intrinsic identifiers
- facts and provenance

Organization
- non-profit
- multi-stakeholder

Software Heritage

As of today, the archive already contains and keeps safe for you the following amount of objects:

- Source files: 8,846,381,610
- Commits: 1,880,663,008
- Projects: 140,348,311
- Directories: 7,506,954,410
- Authors: 38,603,337
- Releases: 15,051,940

Roberto Di Cosmo (CC-BY 4.0) Preserving the Software Commons 18/09/2020 4 / 12
Global development history permanently archived in a unique git-like Merkle DAG

- ~400 TB (uncompressed) blobs, ~20 B nodes, ~280 B edges
So/f_tware 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
Outline

1. Software Source Code: precious part of our Knowledge
2. Meet Software Heritage
3. Demo time!
4. Building for the long term …
5. … and firefighting too!
Outline

1. Software Source Code: precious part of our Knowledge
2. Meet Software Heritage
3. Demo time!
4. Building for the long term …
5. … and firefighting too!
An international, non profit initiative... 

Sharing the vision

United Nations Educational, Scientific and Cultural Organization

Donors, members, sponsors

Platinum sponsors

Gold sponsor

Silver sponsors

Bronze sponsors

And many more ...

www.softwareheritage.org/support/testimonials
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
- infrastructure for researchers
- stepping stone to an international joint effort
UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …

Their call is published on Feb 2019

It’s an important *policy tool*, already referenced and used …

*yes, you can sign it!*

Outline

1. Software Source Code: precious part of our Knowledge
2. Meet Software Heritage
3. Demo time!
4. Building for the long term …
5. … and firefighting too!
It’s urgent to expand the archive

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

Bottomline

*explicit deposit* is important, ...

... and we must promote it...

... but will never be enough.

*(think also of all software dependencies!)*
You can help

Experts are needed ...

- can contribute *listers* and *loaders*? apply for a Sloan funded minigrant!

- knowledgeable in international organizations? UNESCO seeks contributions!

... and everybody can help

- adopt SWH in your work
- save relevant source code
- contribute SWH is free software
- advocate spread the word
Everybody is concerned, everybody can help build

The Library of Alexandria of code
- recover the past
- structure the future

A CERN for Software
- build better software
  - for industry
  - for society as a whole
Appendix
6 SWHIDs by the example

7 News
A word on the trust model for systems of identifiers

Two general classes of systems of identifiers

**intrinsic** computed from the object *(no registry required, fully decentralised)*
(e.g.: chemical notation, music notation, hashes, SWHIDs)

**extrinsic** assigned by an authority *(need a registry)*
(e.g.: passport number, DOI, ARK, RRID, etc.)

See the dedicated blog post for more details
Software Heritage Identifiers (SWHID)

Examples:
- Apollo 11 AGC excerpt,
- Quake III rsqrt

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

Contents

GNU GENERAL PUBLIC LICENSE
Version 3, 29 June 2007

Copyright (C) 2007 Free Software Foundation, Inc. <http://fsf.org>
Everyone is permitted to copy and distribute verbatim copies of
this license document, but changing it is not allowed.

Preamble

The GNU General Public License is a free, copyleft license for
software and other kinds of works.

The licenses for most software and other practical works are designed
To take away your freedom to share and change the works. By contrast,
the GNU General Public License is intended to guarantee your freedom
to share and change all versions of a program-to make sure it remains free
software for all its users. We, the Free Software Foundation, use the
GNU General Public License for most of our software; it applies also to
any other work released this way by its authors. You can apply it to
your programs, too.

When we speak of free software, we are referring to freedom, not
price. Our General Public Licenses are designed to make sure that you
have the freedom to distribute copies of free software (and charge for
them if you wish), that you receive source code or can get it if you
want it, that you can change the software or use pieces of it in new
free programs, and that you know you can do these things.

To protect your rights, we need to prevent others from denying you
these rights or asking you to surrender the rights. Therefore, you
receive a warranty of freedom that allows you to do what you want with
the software.

Preserving the Software Commons 18/09/2020 4 / 6

sha1: 8624bcdae55baeef...
sha256: 8ceb4b9ee5aded...
sha1_git: 94a9ed024d385...
length: 35147
A worked example
A worked example

Directories

```
100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d6a34af6f52cf3cf6b0c2f7bd0648fbd255e77f AUTHORS
100644 blob 94a9ed824d3859793618152ea559a16babbc85e2 LICENSE
100644 blob d9b2665a435a4f38a79a84e8667751d9b95c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5caf2eb4 Makefile
100644 blob 5c7e3a5bbdb0b38682ba7793f40492ed9678bb3 Makefile.local
100644 blob 8617980629c2d4e680404f09aa7496b85b3e67b README.db_testing
100644 blob 76b29f94cf815e086c414d38d78d7ce08ec514e README.dev
040000 tree e1e10ecef948af0b93adb0372a9c89f12e92618a bin
040000 tree 83e56d0beaff7793c77a45a345c80fcb8af503013 debian
040000 tree a34c9c4ba213f0cedc67f9816348d27955577af5 docs
100644 blob f2a6d3c6135a7287bbd76167b81df2ae4f1539 requirements.txt
100755 blob eee147c36caflbcb2d829da8dc026c5b68180c setup.py
040000 tree 224bb4c1f4c67fa01d60947e1abf3 sql
040000 tree 8631c9c77bbbe99310817ab5ba5f1f40c6300be swh
040000 tree 8fb905b56ba8ed692f1209b2773b474c61d66cl utils
```

id: 515f00d44e92c65322aaa9bf3fa097c00dd9c7d
A worked example
Revisions

<table>
<thead>
<tr>
<th>Details</th>
<th>Changes</th>
<th>Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author: Nicolas Dandrimont <a href="mailto:nicolas@dandrimont.eu">nicolas@dandrimont.eu</a></td>
<td>Thu Sep 1 14:26:13 2018</td>
<td></td>
</tr>
<tr>
<td>Committer: Nicolas Dandrimont <a href="mailto:nicolas@dandrimont.eu">nicolas@dandrimont.eu</a></td>
<td>Thu Sep 1 14:26:13 2018</td>
<td></td>
</tr>
<tr>
<td>Subject: provenance.tasks: add the revision -&gt; origin cache task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent: fc3a8b59ca1df4248e6f2c29ab07fee4dc35d10</td>
<td>test...storage: properly pipeline origin and cont...</td>
<td></td>
</tr>
<tr>
<td>provenance.tasks: add the revision -&gt; origin cache task</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
515f00d44e92c65322aaa9bf3fa097c00dd9c7d
tree 515f00d44e92c65322aaa9bf3fa097c00dd9c7d
parent: fc3a8b59ca1df4248e6f2c29ab07fee4dc35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

provenance.tasks: add the revision -> origin cache task
```

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6
A worked example

Releases

object c0c9f16b1e134f593e7567570a1761b156e6eb1d

type commit

tag v0.0.51

tagger Nicolas Dandrimont <nicolas@dandrimont.eu> 1472042163 +0200

Release swh.storage v0.0.51

- Add new metadata column to origin_visit
- Update swh-add-directory script for updated API

--- BEGIN PGP SIGNATURE-----

IQizBAAABCAdBOvQXzTNTFxxuaWNvGzQGRnbfRyaW1vdnQuZXUACgniQ7AWLmo2+
neqpwroadiG50505djxXc+kWJlrYpysS+1kXvEVhJ3WkAxw8K7x2kELD7TSuf
ahpZ/6p3j8lnpe6acC1rXsflcP3L2yftrZeWXXVqRhbxWVMAEo7D3b8qaphw8A5I512
ICB/2jXu639keK6wz1Zkx+8b0jfMY3Sdr6W7Z7K4mUjP0gxyhP75Yo
IcEnWmno7YV7YVls6Cn5d7B7565XqZsAA+becqy6ub7Z2xj+yjUqUcBcyqy3nfr/CL
qs27uxqyzy88YTHGt1yj+iP+XsWbinPs5TTl0oujYbVpK7zhSPFkqjriBh27FkCao
kj6kAWjYkBpMeb+nKvJei.jbrR3+yYWBI3Qo5a1VX0O7h6E1dALCmNepEaKcRKKM

drg/HAXaLS/g0GEdfqlw/5Q7Xk6PvK7P+hsvQjO9itv3GQoQ73nu1npzMc006hi9hAa2zC
Gq/K1P8HF4zvG946wY7zyv0u/ZXGEdusvWU9W7Q2Zk/Win+5Zrzdk6kdrjSjuQOMnRnpTTfUSAXeXhGopkgkXh5YTVp1jPc76USTKDaGe844zZmlk6brOwXCVpYbo
nhh85B5B8MooyF6yT50jU1K7Z0fpYRUGKWdeRGKw6SKxKUOZGxKzy67dUlo0Z9
gwJzhZOFQw7QBD9omAL2+hvPflYdCMeqUq6zgFr+El6VUs=
＝kOxP

-----END PGP SIGNATURE-----

id: 85083a5cc14a441c89dea73f5bd67c3f9c6afdb
A worked example
A worked example

Roberto Di Cosmo

Snapshots

commit 08ffe2b577010952e3eb3e21691466c53a1d9158 refs/heads/atime
commit bd5443a4edf9f3e23b46c3102e4c0abe61c67e2 refs/heads/directory.listing-arrays
commit df96d0a893238569b2f7f0b1c0852723369f9 refs/heads/foo
commit cf7f9eea0a6b22b80405667b5a8610f6f7d680e88 refs/heads/master
commit 7eca197f6c5e6d282a40476e5b1ed9e0b44a361a8f2c refs/heads/tmp/tpo-admin
commit 042a285575d86a095a065247bb35eef4f0252e6e2e refs/heads/tmp/generic-releases
commit 291043b3c1378e7f6b9965997799fd06907c757f55 refs/tags/v0.3.1

tag 72a219913a84ee39996dbb86/7fb800ee7a2eeedc2 refs/tags/v0.3.10

tag 35964eb26e7f8b3e37667f65f2a388bfa4f4f9f5c refs/tags/v0.3.11

tag 33378427a43bb3b43b506a77770e06867f6c0555e refs/tags/v0.3.12

tag 65f7465275b237c5f5903112cbfa036fc3b3b35d refs/tags/v0.3.13

tag 5a6325f3f668d854b588e144286d92a1ee3273bbd refs/tags/v0.3.14

tag 586f8d4b580b475f59307643cb3e477f refs/tags/v0.3.15

tag 8cd8bb885f4998b363177742d289f66e5ee51c refs/tags/v0.3.16

tag a542444ee3f8be33eb5e3287ee635089abc7d6 refs/tags/v0.3.17

tag 22ba13f559b9d12226e5655546e01e96fc4993d9 refs/tags/v0.3.18

tag 66979a4ca8d547f3c0d24aad09dce82536ef47c refs/tags/v0.3.19

tag 32fb5a59f9ca3230aad5f16a503aa32f7a0b37e refs/tags/v0.3.20

tag 3147c3c3c46cfe092780a3e6b0b1237deebdf7c7 refs/tags/v0.3.20

tag 31e8a0d0b11c82be072e76eb6b6073a87908 refs/tags/v0.3.21

tag 3f7b68c287a5ad5b52124257a15d5f85f4ad1df refs/tags/v0.3.22

tag 8cd06de4d73fc5a26789e46b16ac3c72aba4 refs/tags/v0.3.23

...
News: archiving public code

https://code.etalab.gouv.fr
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