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 BANKCALL # SILLY THING AROUND
CADR GOPRF1
TCF GOTOP00M # TERMINATE
TCF P63SPOT3 # PROCEED SEE IF HE’S LYING
```

Quake III source code (excerpt)

```
float q_sqrt(float number)
{
    long l;
    float x2, y;
    const float threehalves = 1.5F;
    x2 = number * 0.5F;
    y = number;
    l = * (long*) &y; // evil floating point bit level hacking
    l = 0x5F3759df - (1 << 1); // what the fuck?
    y = * (float*) &l;
    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.*”

@rdicosmo Roberto Di Cosmo @swheritage
Exponential growth and increasing complexity

Growth of **globally original known content** (source: Rousseau, Di Cosmo, Zacchioli 2019)

- Trend over 20 years: *original content doubles every 22 months* (i.e. ~45% per year)

Complexity of the software ecosystem

- *Millions* of lines of code
- Large *web of dependencies*
  - Easy to break, difficult to maintain
- Decentralised *developer communities*
Precious asset, endangered heritage

We are at a turning point

Software now a critical asset for society, but key scientists/developers are passing away, and source code is getting lost or misplaced while software development skyrockets!

To enable next generation research and software development we need a common, non profit, long term, shared infrastructure that provides

A catalog

An archive

A research infrastructure
Our mission

Collect, preserve and share the source code of all the software that is available

Preserving the past, enhancing the present, preparing the future
A revolutionary infrastructure for software source code

The graph of Software Development
All software development with its history, in a single graph ...

A pillar of Open Science
Reference archive of Research Software

The blockchain of Software Development
... a single Merkle graph, with intrinsic ids for traceability

Reference platform for Big Code
One uniform data structure enables massive machine learning for quality, cybersecurity, etc.
An international, non profit initiative built for the long term

Sharing the vision

And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Platinum sponsors

Gold sponsor

Silver sponsors

Bronze sponsors

On the future of software development  CC-BY 4.0  Apr 7th 2020  7 / 10
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!

Global development history permanently archived in a unique git-like Merkle DAG

- ~400 TB (uncompressed) blobs, ~20 B nodes, ~280 B edges
Questions?

Learn more

Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchiromli
*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 Zacchiromli
*Referencing Source Code Artifacts: a Separate Concern in Software Citation*, Computing in Science & Engineering, 2020, ISSN: 1521-9615

Roberto Di Cosmo, Stefano Zacchiromli
*Software Heritage: Why and How to Preserve Software Source Code*, iPRES 2017
Outline

1. More about Software Heritage
2. The SWH-ID: the source code fingerprint
Highlights from the launching phase

**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
Outline

1. More about Software Heritage
2. The SWH-ID: the source code fingerprint
The SWH-ID schema

```
swh:1:cnt:<41ddb23118f92d7218099a5e7a990cf58f1d07fa
```
The SWH-ID schema

- `schema_version`
- `object_id`
- `prefix`
- `object_type`

- "snp" - snapshot
- "rel" - release
- "rev" - revision
- "dir" - directory
- "cnt" - content
The SWH-ID schema

```
swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa
```

- **schema_version**: This indicates the version of the schema.
- **object_id**: This is the unique identifier for the object.
- **prefix**: This prefix is used to identify the type of the object.
- **object_type**: This specifies the type of the object.

- **lines_ctxt**: Indicates the line numbers of the object.
  - `;lines=64-72`

- **origin_ctxt**: Indicates the origin of the object.
  - `;origin=https://github.com/chrislgarry/Apollo-11`
A worked example
A worked example
Directories

| 100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 | .gitignore |
| 100644 blob 2d6a34afef52cf3cf6b6c2f7bd0648fbd255e77f | AUTHORS |
| 100644 blob 94a9ed824d3859793618152ea559a168bbcb5e2 | LICENSE |
| 100644 blob d9b2665a435a43f8a79a84e8667751dfb895c7bb | MANIFEST.in |
| 100644 blob 524175c2bad0b35b975f79284c2f5a6d5eafe2eb4 | Makefile |
| 100644 blob 5c7e3a5bbeddb038682ba7793f440492ed9678bb3 | Makefile.local |
| 100644 blob 8617980629c2d4e608404f99aa749b085b3e67b | README.db_testing |
| 100644 blob 76b29f94cf815e0869c414d38d787ce08ec514e | README.dev |

```
$ tree
  bin
  debian
  docs
  requirements.txt
  setup.py
  sql
  swh
  utils
```

id: 515f00d44e92c65322aa9bf3fa097c00ddb9c7d
Revisions

SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6
Author: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2018)
Committer: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2018)
Subject: provenance.tasks: add the revision -> origin cache task
Parent: fc3a8b59ca1df424d860f2c29ab07fee4dc35d10 : test...storage: properly pipeline origin and cont...
provenance.tasks: add the revision -> origin cache task

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6
A worked example

Releases

object c0c9f16b1e134f593e7567570a1761b1566eb1d
type commit
tag v0.0.51
tagger Nicola 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 Git SIGNATURE ---

IQ2BAABCAAdBQJXvZTNFnxuxavWHvqFzQGRnbmnfYaW1vonQzZXUACgiphQ7AWLMo2+
neaqywIa650b5Djz8m+4KNWkXjgV5+1Kl1eEHyjWKNwbeXkJtXzXEIDT7Yu
aphZ6pz3q8np6aC1+YpXbRcY3JLYtrrdXeXXXWqXsKwXNMaEo7D868qppwhhBAAS5t2
ICB172jXuxyG913eKFPwv2ZXRg+h8pgMWy35Dr6jW77lZK4UdpGpgy1HP755yo
IGEndWnvo7YV1Vmd3hI5h755mXlaaA+becqdd6ub7Z2xjy+jpUqUCby4qV3nmr/FL
qs2mulkVyYH1IG+1pYV+iU5w8lhnpc3ST1k0yjoylV9PzX1SFp7q1iihi7XZKcaoy
klyi6kAWyU8OBmex++nkV/jei.bR3y+yWBF33q5a1wVoOt374En1DA0CmEnEpEcK6KoKMt
jojIPMax113g0EDfnq67Q6DwPKPf3hgVQLQ3nV3GlQQTJn1q5Mc0206h9rWawzC
GgX11P31h4zoQl5ld6WyyP3zye0U3XG6Wd_wU9vFQ2ZkWjn+c2Mzcdckrj3UQOMn
RpTTUsXrUXeXHGDPgkXsYntv4g0PvC7uST3KDA6e4A4Zm1k0mGmWxXCVpQgYo
nhbIBSHMNOmzvFw75poUb1YK70t1PRUKGwxKGlXKUZCkXz6fj0ij0xq9
kfiw1gZ0MqW2C80omALZ+HvP4kVyc4Mepuh92cYr+EHIvuc8
=KOxP

--- END Git SIGNATURE ---
id: 85083a5cc14a441c89deaa73f5bdf67c3f9c6afdb
A worked example
Snapshots

commit 0f0feb2577010952eb3c3e21691466c53ad9158 refs/heads/atime
commit ba5443a2e4f9fe3e23b46c2792ec4fe0e61c67e ref/heads/directory-listing-arrays
commit dff9eda8f9283e93e85689b27febf1c0562723869c5 refs/heads/foo
commit c7ff79eaeb22b278b94690f75a58e19f672de68e88 refs/heads/master
commit 7eca197ef1e6d20407e54b1ed9e8b44361a8f2c ref/heads/tmpl-config-addr
commit 4b42285f51585058a56d4270b53ee4f3b2352e82e refs/heads/tmpl-generic-releases

tag 2f403c1378c:76896653779f0d6997c757755 refs/tags/v0.1

tag 72a21991a3384e53996d9e6b7bfb0bee57aee2cd refs/tags/v0.2.10

tag 359ab6cbb6d7b3a5337b7879a5e4576308f0239bf8401a06d refs/tags/v0.1.11

tag 33378a43a2a3b2b3a6e777b20d6674f8c92556 refs/tags/v0.1.12

tag 86f7465d275b322c5f90311c2bfa0936cfc3b4a3d refs/tags/v0.1.13

tag 5a6325fe86b645c681e442b2d7d92a1e1e32f7b refs/tags/v0.1.14

tag 5bf6a6a5b8d4575a6b959307643cb3ebac97f refs/tags/v0.1.15

tag 8c9b888c498b2f536377742d28e9f66e05e5c1c refs/tags/v0.1.16

tag a54244639f9be3ce5fba42827ee835b89abc7d6 refs/tags/v0.1.17

tag 228a2f1559d12226e5b535462e1e6fd4c993d9 refs/tags/v0.1.18

tag 669794ac8d549f7cf80244aada0cde825363e4f7c refs/tags/v0.1.19

tag 3cb5af39c2a3f3baadd6519a5382ec275a67 refs/tags/v0.1.20

tag 3147c83358c6c1f64927801e398bb1237e9df2c7 refs/tags/v0.1.20

tag 215e50aba11c8c820727e66eb4b6073a87900 refs/tags/v0.1.21

tag 3fb186c2873a5d2b52124257a1a5d1b5ff280c refs/tags/v0.1.22

tag 8c0bee8dada737f5c542627894e40b16ac3c72baa4 refs/tags/v0.1.23
...

id: b464cad1b66ff266a37b46ea6e7a04b545e904b

A worked example
Demo time

Let’s look at some famous excerpts of source code

### Apollo 11 source code (excerpt)

<table>
<thead>
<tr>
<th>MOVES</th>
<th>Cause</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>P63SPOT3</td>
<td>BIT6</td>
<td># IS THE LR ANTENNA IN POSITION 1 YET</td>
</tr>
<tr>
<td>EXTEND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAND</td>
<td>CHAN33</td>
<td></td>
</tr>
<tr>
<td>EXTEND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BZF</td>
<td>P63SPOT4</td>
<td># BRANCH IF ANTENNA ALREADY IN POSITION 1</td>
</tr>
<tr>
<td>CAF</td>
<td>CODE5050</td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>BANCALL</td>
<td></td>
</tr>
<tr>
<td>CADR</td>
<td>GOPHER1</td>
<td></td>
</tr>
<tr>
<td>TCF</td>
<td>GOTOPO0H</td>
<td># TERMINATE</td>
</tr>
<tr>
<td>TCF</td>
<td>P63SPOT3</td>
<td># PROCEED</td>
</tr>
<tr>
<td>P63SPOT4</td>
<td>TC</td>
<td>BANCALL</td>
</tr>
<tr>
<td>TC</td>
<td>SETPOS1</td>
<td># ENTER</td>
</tr>
<tr>
<td>CADR</td>
<td>BURNBABY</td>
<td></td>
</tr>
</tbody>
</table>

### Quake III source code (excerpt)

```c
float Q_sqrt(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 = 0xF3759df - (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;
}
```

It works!

we have **intrinsic** identifiers for all 20+ billion objects in the archive
The Software Heritage ID schema  (see http://bit.ly/swhpids)

- **swh:1:cnt:**94a9ed024d3859793618152ea559a168bbcb5e2  full text of the GPL3 license
- **swh:1:dir:**d198bc9d7a6bcf6db04f476d29314f157507d505  Darktable source code
- **swh:1:rev:**309cf2674ee7a0749978cf8265ab91a60aea0f7d  a *revision* in the development history of Darktable
- **swh:1:rel:**22ece559cc7cc2364edc5e5593d63ae8bd229f9f  release 2.3.0 of Darktable, dated 24 December 2016
- **swh:1:snp:**c7c108084bc0bf3d81436bf980b46e98bd338453  a *snapshot* of the entire Darktable repository (4 May 2017, GitHub)

Current resolvers: archive.softwareheritage.org and n2t.org