

Software Heritage

joining forces with GitHub to preserve humankind's source code

Roberto Di Cosmo
Director, Software Heritage

Nov 13th, 2019



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

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)

```
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      GOTOP00H      # 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 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.”

Precious asset, *endangered heritage*

We are at a turning point

key people **are passing away** ...

software development **skyrockets** ...

and yet **we are missing** ...

A catalog



A word cloud featuring various software development projects and tools. The most prominent words are 'Git Hub' in large purple letters, 'Sourceforge' in blue, and 'CPAN' in orange. Other visible words include 'Debian', 'Maven', 'Bitbucket', 'GoogleCode', 'Gitlab', 'eTAN', 'CRAN', 'Gitorious', 'BitOs', and 'Adalact'.

An archive



A word cloud with terms related to data loss and disaster. The largest words are 'damage' and 'disaster'. Other significant words include 'deletion', 'malicious', 'obsolete', 'dependencies', 'attack', 'format', 'storage', 'reference', 'dangling', 'wear', 'corruption', 'encryption', 'aging', 'media', and 'tear'.

A research infrastructure



it is **urgent** to take action!



Software Heritage



Our mission

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

Past, present and future

Preserving the past, enhancing the present, preparing the future

Sharing the vision



United Nations
Educational, Scientific and
Cultural Organization



And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors



Platinum sponsors



Golden sponsor



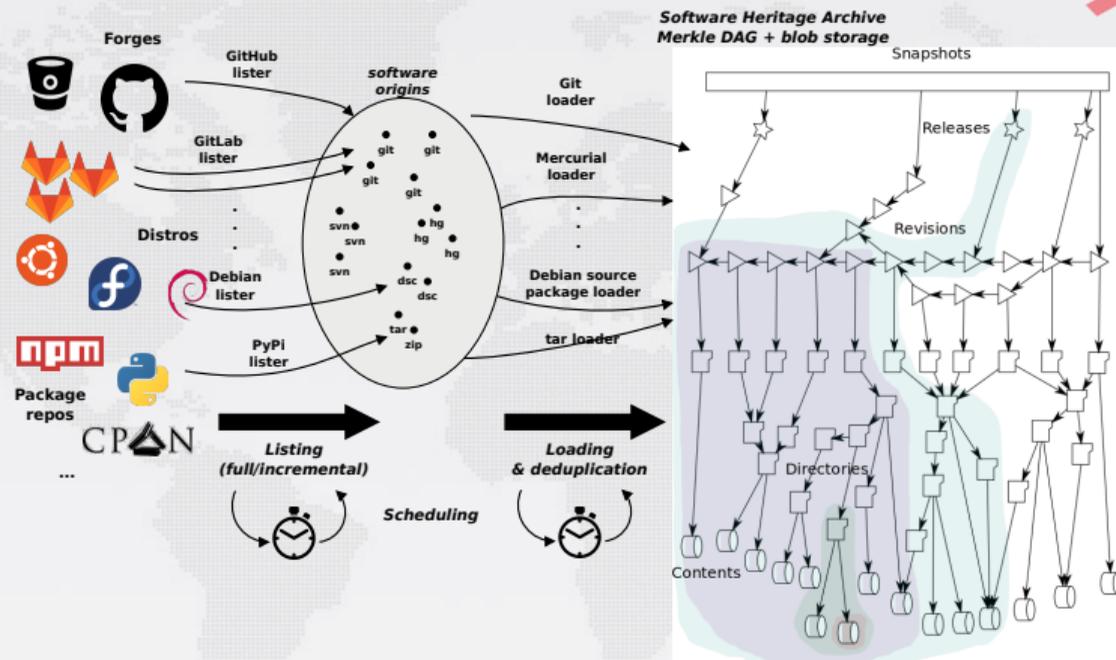
Silver sponsors



Bronze sponsors



A peek under the hood



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

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

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

[We call to] support:

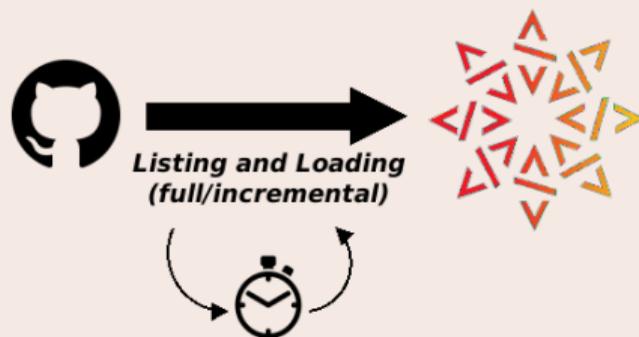
- “efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive”
- “a universal archive, as part of a broad effort at digital preservation, that will ensure persistence of and universal access to software source code”

Support SWH Acquisition Process



<https://www.softwareheritage.org/swhap>

Better GitHub archival in SWH



Join the effort, support the mission!



Software Heritage

www.softwareheritage.org

[@swheritage](https://twitter.com/swheritage)

Library of Alexandria of code

Working together to

- recover the past
- preserve our heritage
- share the knowledge
- prepare the future



Learn more

www.softwareheritage.org/publications



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



Roberto Di Cosmo, Morane Gruenpeter, Stefano Zacchioli
Identifiers for Digital Objects: the Case of Software Source Code Preservation, iPRES 2018: Intl. Conf. on Digital Preservation



Roberto Di Cosmo, Stefano Zacchioli
Software Heritage: Why and How to Preserve Software Source Code, iPRES 2017: Intl. Conf. on Digital Preservation