Software Heritage
Why and How We Preserve all of Mankind’s Source Code

Roberto Di Cosmo
roberto@dicosmo.org
April 20th, 2018
1. Software all around us
2. The Software Heritage initiative
3. Building for the long term
4. Conclusion
Software is everywhere

Software embodies our collective Knowledge and Cultural Heritage
Software Source Code is special

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

“Programs must be written for people to read, and only incidentally for machines to execute.”

Quake 2 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;
}
```

Net. queue in Linux (excerpt)

```c
/*
 * SFB uses two B[i][N] : L x N arrays of bins (L levels, N bins per level)
 * This implementation uses L = 8 and N = 16
 * This permits us to split one 32bit hash (provided per packet by rxhash or
 * external classifier) into 8 subhashes of 4 bits.
 */
#define SFB_BUCKET_SHIFT 4
#define SFB_NUMBUCKETS (1 << SFB_BUCKET_SHIFT) /* N bins per level */
#define SFB_BUCKET_MASK (SFB_NUMBUCKETS - 1)
#define SFB_LEVELS (32 / SFB_BUCKET_SHIFT) /* L */

/* SFB algo uses a virtual queue, named "bin" */
struct sfb_bucket {
    u16 qlen; /* length of virtual queue */
    u16 p_mark; /* marking probability */
};
```

Len Shustek, Computer History Museum

“Source code provides a view into the mind of the designer.”
Apollo 11 Guidance Computer (~60,000 lines), 1969

"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton

Linux Kernel

... now in your pockets!

are we taking care of all this?
Software is fragile
Software lacks its own research infrastructure

Photo: ALMA(ESO/NAOJ/NRAO), R. Hills
1. Software all around us
2. The Software Heritage initiative
3. Building for the long term
4. Conclusion
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
A principled infrastructure


Software Heritage

Technology
- transparency and FOSS
- replicas all the way down

Content
- intrinsic identifiers
- facts and provenance

Organization
- non-profit
- multi-stakeholder
full development history permanently archived
origins: GitHub (automated), Debian (automated), Gitorious, Google Code, GNU
~150Tb raw contents, ~10Tb graph (7+Bn nodes, 60+Bn edges)
Using the archive

Preview just for you: a "wayback machine" for archived code!

- go to http://archive.softwareheritage.org/browse
- use the credentials: devoxx / 2018

No time for a demo, let’s highlight some features...

Origin search

Directory browsing

Revisions as diffs
1. Software all around us
2. The Software Heritage initiative
3. Building for the long term
4. Conclusion
Growing Support

Landmark Inria Unesco agreement, April 3rd, 2017

Sharing the vision

Contributing to the mission

Roberto Di Cosmo www.dicosmo.org

Roberto Di Cosmo www.dicosmo.org
You can help!

Coding, advice see forge.softwareheritage.org

Current development priorities … all contributions equally welcome!

★★★★ documentation
★★ listers/loaders for unsupported forges, VCS
★★ Web UI improvements

Yes, we’ll be hiring, see www.softwareheritage.org/jobs

Funding
• make your company a sponsor: sponsorship.softwareheritage.org
• give your own contribution: www.softwareheritage.org/donate

Spread the word!
• follow and relay project news
• share the vision, tell others how to support the mission

Roberto Di Cosmo www.dicosmo.org
Outline

1. Software all around us
2. The Software Heritage initiative
3. Building for the long term
4. Conclusion
Software Heritage

www.softwareheritage.org  @swheritage

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