Software Heritage
Why and How We Are Building the Universal Software Knowledge Base

Roberto Di Cosmo
roberto@dicosmo.org

October 26th, 2017
1. Software all around us
2. The Software Heritage initiative
3. Status
4. Building for the long term
5. Conclusion
Software is everywhere
Software Source Code is special

Harold Abelson, Structure and Interpretation of Computer Programs

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

Quake 2 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;
}
```

Net. queue in Linux (excerpt)

```c
/*
 * SFB uses two B[][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 ((16 / SFB_BUCKET_SHIFT) / L */

/* SFB algo uses a virtual queue, named "bLn" */
struct sfh_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.”
"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton
Software is spread all around
Software is fragile
Software lacks its own research infrastructure

Photo: ALMA(ESO/NAOJ/NRAO), R. Hills
Looking at the past
- a lot of old software misplaced, lost, or behind barriers, but…
- most founding fathers are still here, and willing to share
- urgent to collect their knowledge

Only a few years left.

Looking at the future
- software development and use skyrocket: more programmers, and more code!
- essential to provide a universal platform for all the future software source code

Every year that goes by makes the problem worse.

it is urgent to take action!
1. Software all around us

2. The Software Heritage initiative

3. Status

4. Building for the long term

5. Conclusion
The Software Heritage Project

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
Preserving and sharing the world’s software heritage

A structured archive of all of the world’s software

- preserve humanity’s technological and scientific knowledge
- enable continued access to all digital documents and information
- precious resource for education
- WikiPedia of software
Universal knowledge base for all industrial software components

- a single entry point to discover, explore and reuse source code
- simplifies traceability (licenses, vulnerabilities, supply chain management)

Reference platform for Big Code

- unique point of observation for all software development
- big data, machine learning paradise: trends, coding patterns, code completion…
Supporting more accessible and reproducible science

A global library referencing all software used in all research fields

- completes the infrastructure for Open Access in science
- provides intrinsic persistent identifiers needed for scientific reproducibility
- enables large scale, verifiable software studies
Outline

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

~150 TB blobs, ~5 TB database (as a graph: ~7 B nodes + ~60 B edges)

Our sources

- GitHub — full, up-to-date mirror
- Debian — automation in progress done; GNU
- Gitorious, Google Code — processing (Archive Team & Google)
- Bitbucket, FusionForge(s) — WIP

The richest source code archive already, … and growing daily!
Much more than an archive!

Merkle tree (R. C. Merkle, Crypto 1979)

Combination of
- tree
- hash function

Classical cryptographic construction
- fast, parallel signature of large data structures
- widely used (e.g., Git, blockchains, IPFS, ...)
- built-in deduplication
Using the archive

Features...

- (done) lookup by content hash
- browsing: "wayback machine" for archived code
  - (done) http://archive.softwareheritage.org/api
  - (in progress) via Web UI
- (in progress) download: wget / git clone from the archive
- (in progress) deposit of source code bundles directly to the archive
- (todo) provenance lookup for all archived content
- (todo) full-text search on all archived source code files

... and much more than one could possibly imagine

all the world’s software development history in a single graph!
You can help!

**Coding**
- forge.softwareheritage.org — our own code
- sympa.inria.fr/sympa/info/swh-devel — devel mailing list

**Current development priorities**

- ★★★ listers for unsupported forges, distros, pkg. managers
- ★★★ loaders for unsupported VCS, source package formats

... all contributions equally welcome!

See the dedicated www.softwareheritage.org/community/blog entry

**We are trying to make it easy**
- forge.softwareheritage.org/tag/easy_hack/
Outline

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

Cultural Heritage  Industry  Research  Education

Software Heritage

Open approach
- Transparency
- Free Software
- User and contributor community building

Objectiveness
- Facts and provenance
- *Intrinsic* identifiers
- Full development history

Long term
- Multi-stakeholder
- Nonprofit
- Replication *at all layers*
Sponsoring Software Heritage work

Microsoft

>= 100Ke/year

>= 50Ke/year

>= 25Ke/year

>= 10Ke/year

Intel

Societe Generale

Huawei

DANS

Nokia Bell Labs

GitHub

Data Archiving and Networked Services

Universita di Bologna
Sharing the Software Heritage vision

See more

http://www.softwareheritage.org/support/testimonials
April 3rd, 2017: landmark Inria Unesco agreement…

September 28th, 2017: Mauritius Call on information access

https://www.softwareheritage.org/blog
Outline

1. Software all around us
2. The Software Heritage initiative
3. Status
4. Building for the long term
5. Conclusion
An unique opportunity

Library of Alexandria of code

Take *urgent* action to
- recover the past
  - founding fathers still here
- structure the future
  - programming skyrockets

A CERN for Software

Build a *common infrastructure*
- supporting industry needs
- enabling software research
- fostering better science
- for society as a whole

Photo: ALMA(ESO/NAOJ/NRAO), R. Hills
Come in, we’re open!

Software Heritage

www.softwareheritage.org

@swheritage

Everybody is needed, and welcome!

partnerships, mirrors

funding

our own code

jobs / internships

mailto:roberto@dicosmo.org

sponsorship.softwareheritage.org

forge.softwareheritage.org

www.softwareheritage.org/jobs

EclipseCon Europe 2017

Evaluate the Sessions
Sign in and vote at eclipsecon.org

-1 0 +1
All the source code
Online, open source code: automation overview
All the source code, strategies
6 Research challenges
Selected research challenges: building the archive

- Distributed infrastructure
  - efficient p2p protocols
- Metadata
  - 65M projects
- Data compression
- Software phylognetics
Selected research challenges: using the archive

- Project classification
  - 65M projects

- Code search
  - AST, functions, features

- Software as Big Data
- Efficient data representation