Software Heritage

Building the Universal Software Archive

Roberto Di Cosmo

roberto@dicosmo.org

November 29th, 2016
Open Source Forum
Société Générale

Software Heritage

Short Bio

Roberto Di Cosmo
Computer Science professor in Paris
now working at INRIA
20 years of Free and Open Source Software

1999 *DemoLinux* – first live GNU/Linux distro

2007 *Free Software Thematic Group*
150 members 40 projects 200Me

2008 *Mancoosi project* `www.mancoosi.org`

2010 *IRILL* `www.irill.org`

2015 *Software Heritage* at INRIA

Software is Pervasive

At the heart of *our society*

- communication, entertainment
- administration, finance
- health, energy, transportation
- education, research, politics
- …

At the heart of *technology*

- house appliances ≈ 10M SLOC
- phones ≈ 20M SLOC, *cars* ≈ 100M SLOC
- Internet of things, …

Software is Knowledge

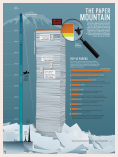
Key mediator for accessing *all* information (c) Banski

Information is a main pillar of our modern societies.

> *Absent an ability to correctly interpret digital information, we are left with […] "rotting bits" […] of no value.*
>
> *Vinton G. Cerf IEEE 2011*

Software is *an essential component* of modern scientific research

> *[…] the vast majority describe experimental methods or sofware that have become essential in their fields.*
>
> Top 100 papers (Nature, October 2014)



Bottomline: Sofware embodies our *Knowledge* and *Cultural Heritage*

It *must* be *collected*, *referenced* and made *accessible*!

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Our mission

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

Past, present and future

Preserving the past, *enhancing* the present, *preparing* the future.

The source code matters!

"The source code for a work means the preferred form of the work for making modifications to it."

GPL Licence

Hello World

Program (excerpt of binary)

```
4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
4004f9: 90
4004fa: 5d
4004fb: c3
```

Program (source code)

```
/* Hello World program */

#include<stdio.h>

void main()
{
    printf("Hello World");
}
```

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)

```
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;
}
```

Network queue in Linux (excerpt)

```
/*
 * SFB uses two B[l][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."

The reference repository of all Source Code

Fashion victims

- many disparate development platforms
- a myriad places where distribution may happen
- projects tend to migrate from one place to the other over time

One place to bind them…

… in Software Heritage you can find and search *all* the source code

The universal archive of Source Code

like all digital information, Software Source Code is fragile

- inconsiderate and/or malicious code loss (e.g., Code Spaces)
- business-driven code loss (e.g., Gitorious, Google Code)
- for obsolete code: physical media decay (data rot)

If (a repository on) GitHub goes away …

… Software Heritage will have a copy of it!

Building the Software Research Infrastructure

Photo: ALMA(ESO/NAOJ/NRAO), R. Hills

A wealth of software research on crucial issues…

- safety, security; test, verification, proof;
- software engineering, software evolution;
- empirical and big data studies;

If you study the stars, you go to Atacama…

… Software Heritage is the *very large telescope* of source code

Better software for industry and society

A unique reference catalog of all industrial software components

- a single entry point to discover, explore and reuse source code
- eases vulnerability tracking for more secure software
- simplifies traceability for better software integration
- ensures long term preservation of critical software

The people

The core team

- Roberto Di Cosmo
- Stefano Zacchiroli
- Nicolas Dandrimont (Engineer)
- Antoine Dumont (Engineer)
- and *Jordi, Quentin and Guillaume*

Scientific advisors

- Serge Abiteboul (French Science Academy)
- Jean-François Abramatic (former W3C director)
- Gerard Berry (CNRS Gold Medal, French Science Academy)
- Julia Lawall (Coccinelle, Linux Kernel, Outreachy)

The archive

Our sources

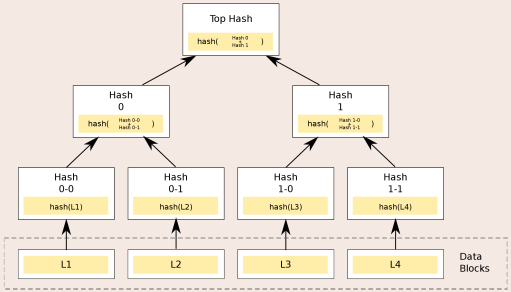
- GitHub — all public repositories as of August 2016
- Debian — daily snapshots of all suites since 2005–2015
- GNU — all releases as of August 2015
- Gitorious — retrieved full mirror from Archive Team
- Google Code — retrieved full mirror from Google

Some numbers

The *richest* source code archive already, … and growing daily!

Merkle trees

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



Combination of

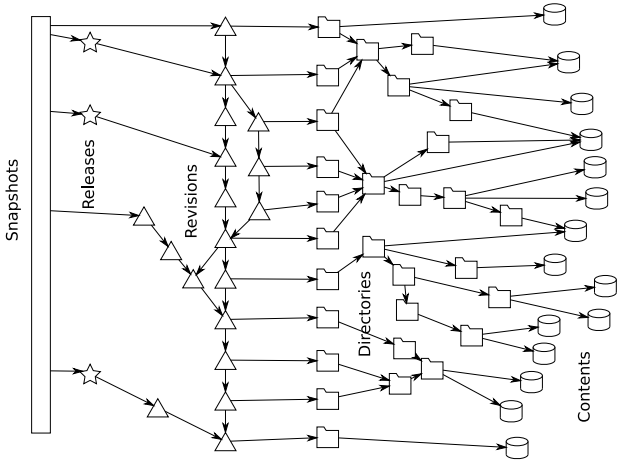
- tree
- hash function

Classical cryptographic construction

- fast, parallel signature of large data structures
- widely used by *Git*, *Bitcoin*, etc.
- natural extension: Merkle *DAG*

The archive in a few pictures

A giant (extended) Merkle DAG



The road ahead

Planned features…

- *lookup* by content hash (done)
- *download*: wget and git clone from Software Heritage
- *provenance information* for all archived code and metadata
- *browsing*: wayback machine for archived code and its history
- *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!

that makes a 150TB archive / 5TB database already…

An ambitious, worldwide initiative

Inria as initiator

- founding partner of the W3C,
- creating a non profit, international organisation

Software Heritage benefits society as a whole

- agreement to be signed in the presence of the highest dignitaries
  - preservation of knowledge embedded in software
  - access to the knowledge embedded in software

Support and *first partners*

ACM, Nokia Bell Labs, Creative Commons, DANS, Eclipse, Engineering, FSF, OSI, GitHub, GitLab, IEEE, Informatics Europe, Microsoft, OIN, OW2, SIF, SFC, SFLC, The Document Foundation, The Linux Foundation, …

Conclusion

Software Heritage is

- a revolutionary *reference archive* of *all* software ever written
- a fantastic new tool for *research* and *industry*
- an international, open, nonprofit, *mutualized infrastructure*
- at the service of our society, at the service of mankind!

Now open

`www.softwareheritage.org` - *sponsoring, partnerships*
`wiki.softwareheritage.org` - *working groups, leads*
`forge.softwareheritage.org` - *our own code*

Questions?