Software Heritage

Building an essential facility for the digital age

Roberto Di Cosmo
Inria and University Paris Diderot

roberto@dicosmo.org

October 24th 2017
ECSS 2017

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Software is everywhere

Software embodies our collective Knowledge and Cultural Heritage

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)

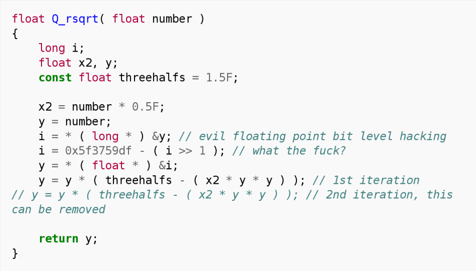
```
/* Hello World program */

#include<stdio.h>

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

Source code is essential

Harold Abelson, Structure and Interpretation of Computer Programs

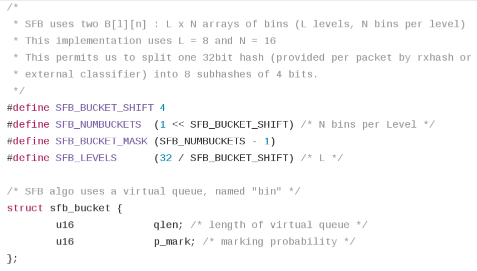
“Programs must be written for people to read, and only incidentally for machines to execute.”

Quake 2 source code (excerpt)



```c
float Q_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;
}
```

Net. queue in Linux (excerpt)



```c
/*
 * 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.”

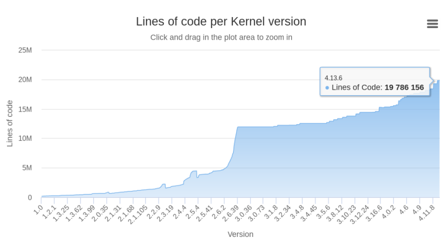
~ 50 years, a lightning fast growth

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



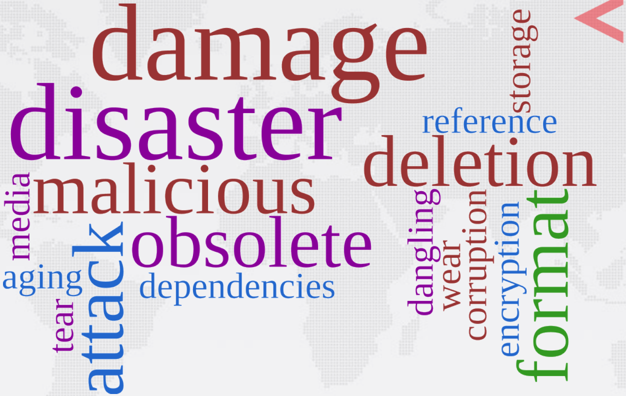
… now in your pockets!

are we taking care of all this?

Software is spread all around

Sourceforge
Debian
CPAN
Gitorious
Maven
Inria
Bitbucket
GitHub
BerliOs
CTAN
Gitlab
CRAN
GoogleCode
Adullact

Software is fragile



Software lacks its own research infrastructure

A wealth of software research on crucial issues...

- safety, security, test, verification, proof
- software engineering, software evolution
- big data, machine learning, empirical studies

If you study the stars, you go to Atacama...

... where is the *very large telescope* of source code?

We are at a turning point

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 skyrockets: 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!

The Software Heritage Project

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Our mission

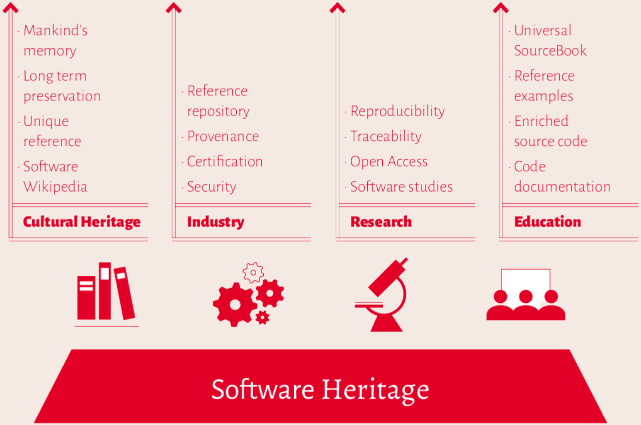
Collect, preserve and share the *source code* of *all the software* that is publicly available.

Past, present and future

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

We are working on the foundations

One infrastructure to build them all



- Mankind's memory
- Long term preservation
- Unique reference
- Software Wikipedia

Cultural Heritage

- Reference repository
- Provenance
- Certification
- Security

Industry

- Reproducibility
- Traceability
- Open Access
- Software studies

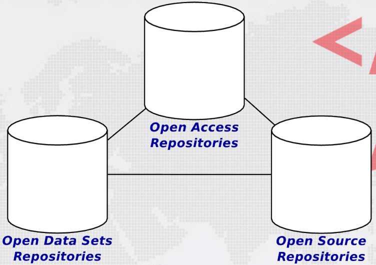
Research

- Universal SourceBook
- Reference examples
- Enriched source code
- Code documentation

Education

Software Heritage

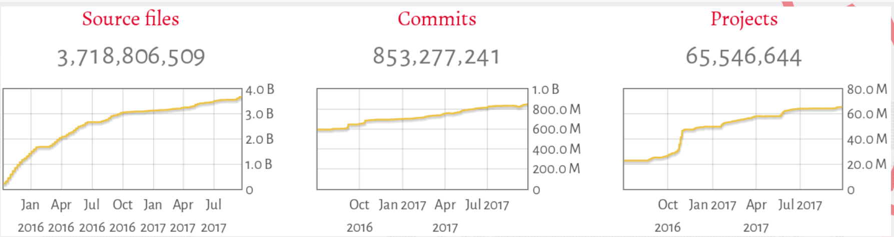
Supporting more accessible and reproducible science



A global library referencing all software used in all research fields

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

Archive coverage



Source files

3,718,806,509

Commits

853,277,241

Projects

65,546,644

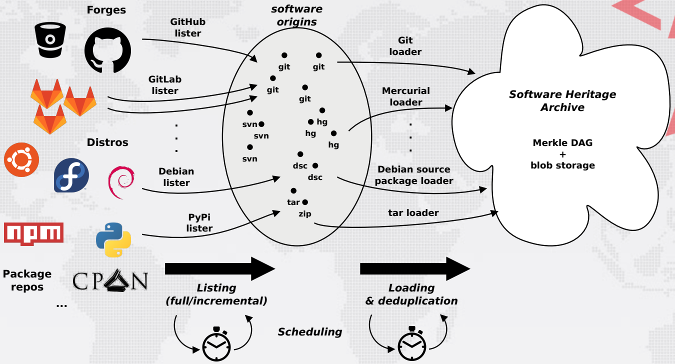
~150 TB blobs, ~5 TB database (as a graph: ~7 B nodes + ~60 B edges)

Our sources

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

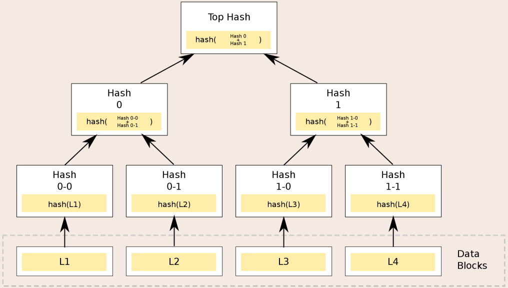
The *richest* source code archive already, … and growing daily!

A complex task



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

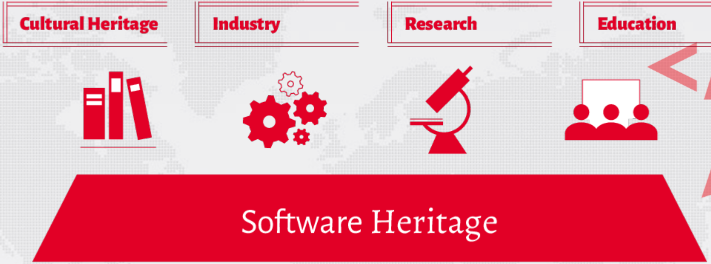
Features…

- (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!

Our principles



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*

Three pillars

Science and technology

- build on sound basis
- fantastic playground for research

Resources

- fund the effort
- transfer to industry and society

Awareness

- promote public and private policies
- community building

Selected research challenges

Building the archive

- data compression
- metadata alignment
- distributed infrastructure
- software phylogenetics
- ...

Using the archive

- project classification
- code search
- efficient (big) data representation
- visualization
- ...

... ethical and legal issues too ...

doors are wide open for collaboration!

Sponsoring Software Heritage work

Inria
INVENTEURS DU MONDE NUMÉRIQUE

>= 100Ke/year	Microsoft	intel®	SOCIETE GENERALE
>= 50Ke/year			
>= 25Ke/year		HUAWEI	
>= 10Ke/year	Data Archiving and Networked Services DANS	NOKIA Bell Labs	ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA DIPARTIMENTO DI INFORMATICA - SCIENZA E INGEGNERIA
			GitHub

Sharing the Software Heritage vision

IEEE Software
GitLab
ENGINEERING
eclipse
OW2
fsfe
nlnet FOUNDATION
SIGSOFT SPECIAL INTEREST GROUP ON SOFTWARE ENGINEERING
Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
INAPI
Gobierno de Chile
THE LINUX FOUNDATION
ADULLACT
INFORMATICS EUROPE
Computer History Museum
software freedom conservancy
Software Freedom Law Center
AdaCore The GNAT Pro Company
gandi.net
TDF
FREE SOFTWARE FOUNDATION
creative commons
Société Informatique de France
open source initiative
openinventionnetwork

See more

http://www.softwareheritage.org/support/testimonials

Going global

April 3rd, 2017: landmark Inria Unesco agreement…

https://www.softwareheritage.org/blog

September 28th, 2017

September 2017: Mauritius Call on information access

Going global

April 3rd, 2017: landmark Inria Unesco agreement…

https://www.softwareheritage.org/blog

September 28th, 2017

Mauritius Call on information access

Forthcoming: Declaration on Software Relevance, Preservation and Access

An unique opportunity for Computer Science

The History of Computing

Take *urgent* action to

- recover the past
  - founding fathers still here
- structure the future
  - programming skyrockets

A CERN for CS

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

Build a common infrastructure

- for research on programming
- supporting all researchers
- helping industry
- for society as a whole

Getting involved

Voice

- testimonials.softwareheritage.org
- contribute to the declaration
- help reach out to industry

Knowledge

- science
- ethics

Network

- joint research projects
- create a Software Heritage mirror

Zoom on the mirror network

Setting up a mirror at your institution

A double advantage!

- Contribute to the global mission
  - replicate the data
  - lower the risk of loss
  - increase access bandwidth
- Increase local visibility and use
  - access to a unique data set for your research
  - leverage the Software Heritage global outreach
  - increase local authorities support for CS

Come in, we're open!

Questions?

learn more

social	@swheritage
main website	www.softwareheritage.org
sponsoring / partnership	sponsorship.softwareheritage.org
talks/press/dataset	annex.softwareheritage.org
our own code	forge.softwareheritage.org