Software Heritage

Why and How to Preserve Software Source Code

Roberto Di Cosmo, **Stefano Zacchiroli**

University Paris Diderot & Inria — zack@upsilon.cc

27 September 2017
iPRES 2017 — 14th International Conference on Digital Preservation
Kyoto, Japan

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Software is everywhere

Source code matters!

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

— GPL Licence

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

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
```

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");
}
```

Our Software Commons

Definition (Commons)

The commons is the cultural and natural resources accessible to all members of a society, including natural materials such as air, water, and a habitable earth. These resources are held in common, not owned privately. https://en.wikipedia.org/wiki/Commons

Definition (Software Commons)

The software commons consists of all computer software which is available at little or no cost and which can be altered and reused with few restrictions. Thus *all open source software and all free software are part of the [software] commons.* [...]

https://en.wikipedia.org/wiki/Software_Commons

Our Software Commons

Definition (Commons)

The commons is the cultural and natural resources accessible to all members of a society, including natural materials such as air, water, and a habitable earth. These resources are held in common, not owned privately. https://en.wikipedia.org/wiki/Commons

Definition (Software Commons)

The software commons consists of all computer software which is available at little or no cost and which can be altered and reused with few restrictions. Thus *all open source software and all free software are part of the [software] commons.* [...]

https://en.wikipedia.org/wiki/Software_Commons

Source code is *a precious part* of our commons

are we taking care of it?

Software is spread all around

Fashion victims

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

Software is spread all around

Fashion victims

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

Where is the place …

where we can find, track and search *all* source code?

Software is fragile

Like all digital information, FOSS 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)

Software is fragile

Like all digital information, FOSS 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)

Where is the archive…

where we go if (a repository on) GitHub or GitLab.com goes away?

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

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?

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Our mission

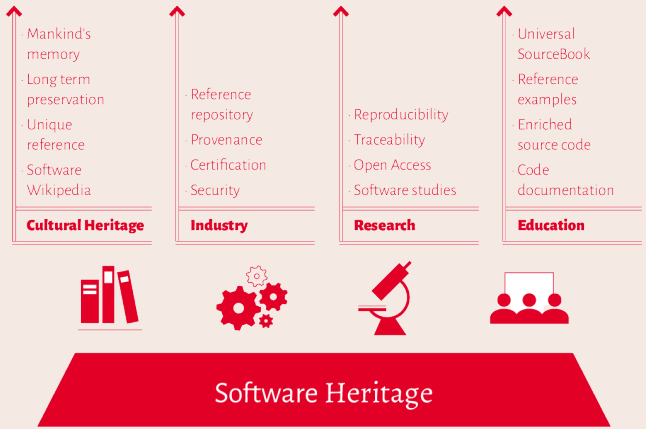
Collect, preserve and share the *source code* of *all the software* that is publicly available.

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



Core principles

Technology

- transparency and FOSS
- replicas all the way down

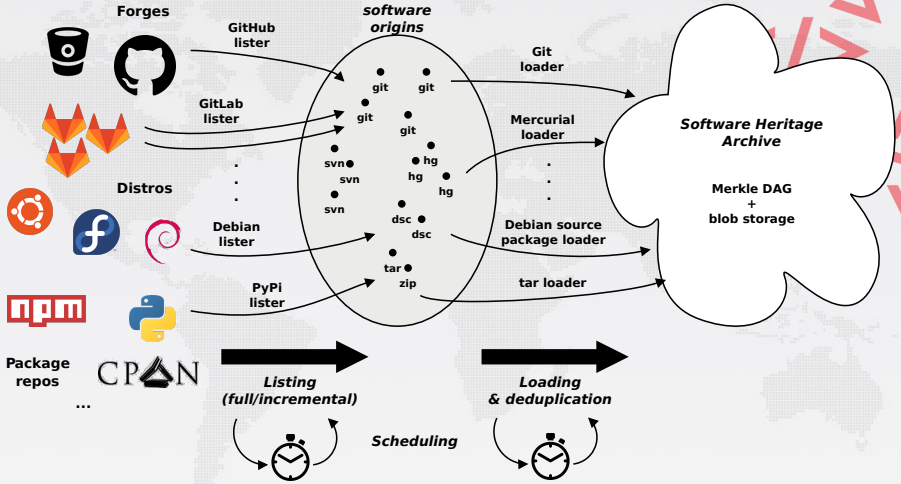
Content

- source code first
- no a priori selection
- intrinsic identifiers
- facts and provenance

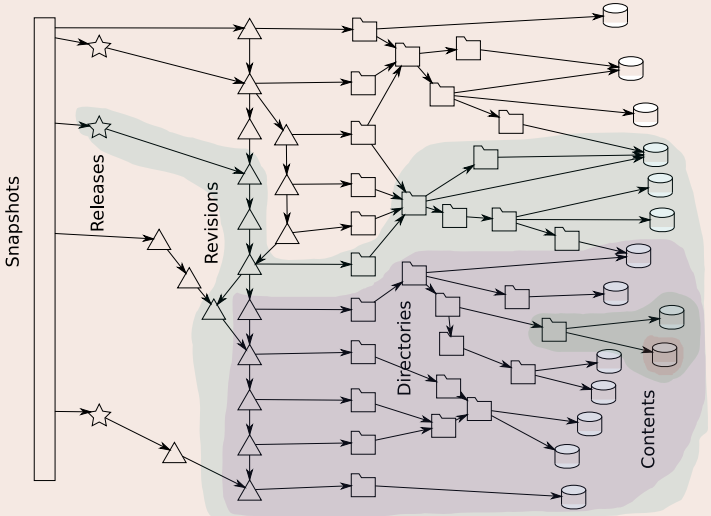
Organization

- non-profit
- multi-stakeholder

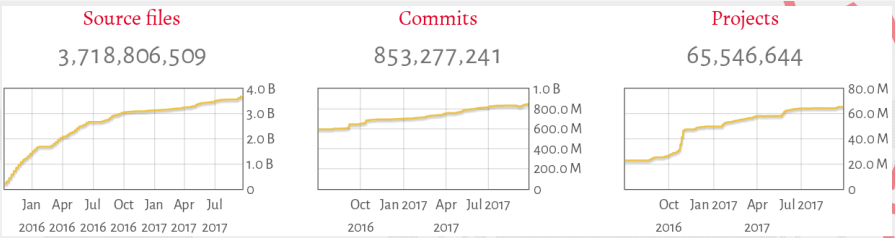
Data flow


Forges
GitHub lister
GitLab lister
Distros
Debian lister
PyPi lister
Package repos
CPAN
...
software origins
git
svn
hg
dsc
tar
zip
Git loader
Mercurial loader
Debian source package loader
tar loader
Software Heritage Archive
Merkle DAG + blob storage
Listing (full/incremental)
Scheduling
Loading & deduplication

The archive: a (giant) Merkle DAG



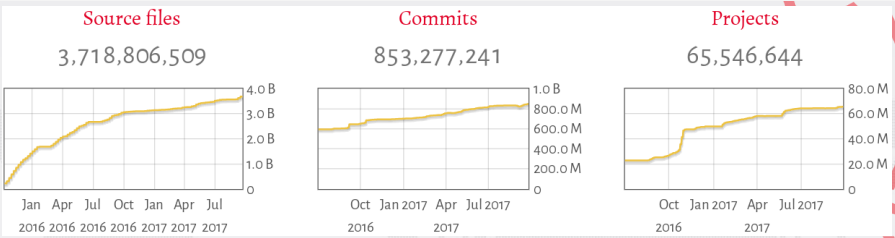
Archive coverage



Our current sources

- GitHub
- Debian, GNU
- WIP: Gitorious, Google Code, Bitbucket

Archive coverage

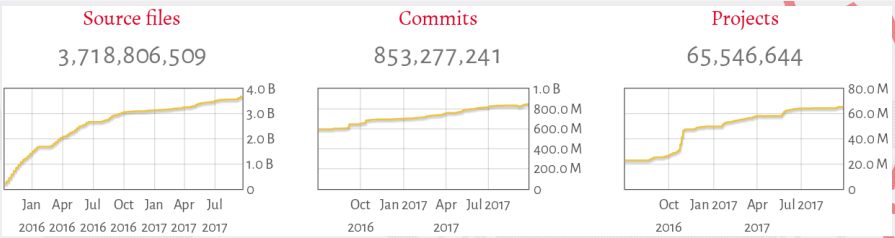


Our current sources

- GitHub
- Debian, GNU
- WIP: Gitorious, Google Code, Bitbucket

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

Archive coverage



Our current sources

- GitHub
- Debian, GNU
- WIP: Gitorious, Google Code, Bitbucket

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

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

Roadmap

Features...

- (done) lookup by content hash
- browsing: "wayback machine" for archived code
  - (done) via Web API
  - (todo) via Web UI
- (todo) download: `wget` / `git clone` from the archive
- (todo) 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

Roadmap

Features…

- (done) lookup by content hash
- browsing: "wayback machine" for archived code
  - (done) via Web API
  - (todo) via Web UI
- (todo) download: `wget` / `git clone` from the archive
- (todo) 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!

The Software Heritage community

Sponsors

Inria — INVENTEURS DU MONDE NUMÉRIQUE

Microsoft · intel · SOCIETE GENERALE · HUAWEI · DANS · NOKIA Bell Labs

Testimonials

Software · GitLab · eclipse · CW2 · fsfe · nlnet FOUNDATION · SIGSOFT · ENGINEERING · THE LINUX FOUNDATION · ADULLACT · INFORMATICS EUROPE · IDF · GitHub · software freedom conservancy · Software Freedom Law Center · Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE · FREE SOFTWARE FOUNDATION · creative commons · AdaCore The GNAT Pro Company · open source initiative · openinventionnetwork

UNESCO/Inria agreement (April 3rd, 2017)

UNESCO

Conclusion

- it is now urgent to preserve software source code
- Software Heritage is taking a systematic approach at it
- Software Heritage has already assembled the largest archive to date
- Software Heritage has synergies with cultural, research, and industry needs
- it is a shared infrastructure that can benefit us all…
- … we should collaborate and pool resources to make it so

Come in, we're open!

`www.softwareheritage.org` — learn more
`www.softwareheritage.org/support/sponsors/` — sponsoring info
`wiki.softwareheritage.org` — working groups
`forge.softwareheritage.org` — our own code

Questions?