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

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

— GPL Licence

Hello World

/* Hello World program */
#include<stdio.h>
void main()
{
  printf("Hello World");
}
"The source code for a work means the preferred form of the work for making modifications to it."

— GPL Licence
"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)

<table>
<thead>
<tr>
<th>Address</th>
<th>Hex</th>
<th>Decimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>4004e6</td>
<td>55</td>
<td>85</td>
</tr>
<tr>
<td>4004e7</td>
<td>c8 e9</td>
<td>192</td>
</tr>
<tr>
<td>4004ea</td>
<td>b9 8a</td>
<td>2736</td>
</tr>
<tr>
<td>4004ef</td>
<td>b8 00</td>
<td>200</td>
</tr>
<tr>
<td>4004f4</td>
<td>c8 c7</td>
<td>279</td>
</tr>
<tr>
<td>4004f9</td>
<td>f8 fe</td>
<td>586765</td>
</tr>
<tr>
<td>4004fa</td>
<td>5d</td>
<td>93</td>
</tr>
<tr>
<td>4004fb</td>
<td>c3</td>
<td>195</td>
</tr>
</tbody>
</table>

Program (source code)

```c
/* Hello World program */

#include<stdio.h>

void main()
{
    printf("Hello World");
}
```
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](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](https://en.wikipedia.org/wiki/Software_Commons)
<table>
<thead>
<tr>
<th>Definition (Commons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The <strong>commons</strong> 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. <a href="https://en.wikipedia.org/wiki/Commons">https://en.wikipedia.org/wiki/Commons</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Definition (Software Commons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The <strong>software commons</strong> consists of all computer software which is available at little or no cost and which can be altered and reused with few restrictions. Thus <em>all open source software and all free software</em> are part of the [software] commons. […] <a href="https://en.wikipedia.org/wiki/Software_Commons">https://en.wikipedia.org/wiki/Software_Commons</a></td>
</tr>
</tbody>
</table>

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
So/ftware 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?
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)
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?
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
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

Debian lister

PyPi lister

Package repos

Listing (full/incremental)

Loading & deduplication

Scheduling

Software Heritage Archive

Merkle DAG + blob storage

Git loader

Mercurial loader

Debian source package loader

tar loader

dsc

git

git

git

svn

svn

svn

svn

dsc

dsc

tar

zip

...
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

...and much more than one could possibly imagine

all the world's sofware development history in a single graph!

Stefano Zacchirolı
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

Testimonials

UNESCO/Inria agreement (April 3rd, 2017)
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?