Software Heritage
Building the Universal Software Archive

Stefano Zacchiroli
Inria & University Paris Diderot
zack@softwareheritage.org

21 September 2016
OW2con
Paris, France
So/f_tware is pervasive

At the heart of our society

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

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 knowledge

**Key mediator for accessing all information**

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

**Essential component of modern scientific research**

[...] the vast majority describe experimental methods or software that have become essential in their fields.

Top 100 papers (Nature, October 2014)
Software is knowledge

Key mediator for accessing all information

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

Essential component of modern scientific research

[...] the vast majority describe experimental methods or software that have become essential in their fields.

Top 100 papers (Nature, October 2014)

Bottom line: software embodies our Knowledge and Cultural Heritage

It must be collected, preserved, referenced and made accessible!
So/f_tware 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

- 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 website disappears you go to the Internet Archive…

… where do you go if (a repository on) GitHub goes away?
Mission

Collect, organise, preserve and share all the software source code that lies at the heart of our culture and our society.

https://www.softwareheritage.org/
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
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

<table>
<thead>
<tr>
<th>Source files</th>
<th>Commits</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,970,266,880</td>
<td>644,628,800</td>
<td>25,258,776</td>
</tr>
</tbody>
</table>

The richest source code archive already, ... and growing daily!

Stefano Zacchirolı
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

<table>
<thead>
<tr>
<th>Source files</th>
<th>Commits</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,970,266,880</td>
<td>644,628,800</td>
<td>25,258,776</td>
</tr>
</tbody>
</table>

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

- *lookup* by hashes for contents (done)
- *download*: git clone from Software Heritage
- *provenance information* for all the content
- *browsing*: wayback machine for software source code
- *full text search*: dive into the Software Heritage archive
The road ahead

Planned features...

- **lookup** by hashes for contents (done)
- **download**: git clone from Software Heritage
- **provenance information** for all the content
- **browsing**: wayback machine for software source code
- **full text search**: dive into the Software Heritage archive

... and much more one could possibly imagine

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

*that makes a 150TB archive / 5TB database already...*
Inria as initiator

- funds the *bootstrap phase* of Software Heritage
- an agreement with UNESCO is coming soon!
Making it happen

Inria as initiator

- funds the *bootstrap phase* of Software Heritage
- an agreement with UNESCO is coming soon!

Testimonials and early partners

ACM, 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, ...

Going global

building an *open, multistakeholder, nonprofit* organisation
Software Heritage is

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

Software Heritage is

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

Now open

www.softwareheritage.org — sponsoring, partnerships
wiki.softwareheritage.org — working groups, leads
forge.softwareheritage.org — our own code

Questions?