Our Software Heritage

Goal and Enabler for Digital Preservation

Stefano Zacchiroli
Software Heritage — zack@upsilon.cc, @zacchiro

29 November 2018
Digital cultural heritage preservation
Deutsche Nationalbibliothek — Frankfurt, Germany

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Software is everywhere

Software is everywhere

Software embodies a growing part of…

… our scientific, technical, and cultural heritage

Source code: executable and human readable knowledge

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

GPL Licence

Source code: executable and human readable knowledge

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

GPL Licence

```
Hello World
```

Source code: executable and human readable knowledge

"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: executable and human readable knowledge

"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: enabler for all digital preservation

Castagné, M. (2013). *Consider the source: The value of source code to digital preservation strategies.* SLIS Student Research Journal, 2(2)

Source code: enabler for all digital preservation

Castagné, M. (2013). *Consider the source: The value of source code to digital preservation strategies.* SLIS Student Research Journal, 2(2)

- software mediates our access to all sorts of data — music, photos, games, etc.
- software rot destroys our ability to access such data
- state-of-the-art mitigation techniques: emulation, open standards

Source code: enabler for all digital preservation

Castagné, M. (2013). *Consider the source: The value of source code to digital preservation strategies.* SLIS Student Research Journal, 2(2)

- software mediates our access to all sorts of data — music, photos, games, etc.
- software rot destroys our ability to access such data
- state-of-the-art mitigation techniques: emulation, open standards

- software source code preservation is the end game, our last resort if/when everything else fails
- use cases:
  - rebuilding software from source
  - extracting knowledge for clean slate implementation

~ 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

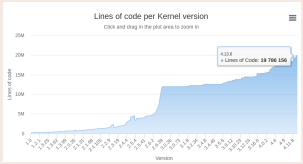
~ 50 years, a lightning fast growth

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!

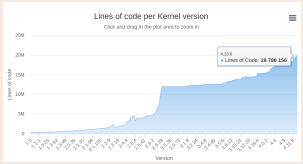
~ 50 years, a lightning fast growth

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!

We are now at a turning point in the history of software technology: are we taking care of all this?

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.

Archiving goals

Targets: VCS repositories & source code releases (e.g., tarballs)

We DO archive

- file content (= blobs)
- revisions (= commits), with full metadata
- releases (= tags), ditto
- where (origin) & when (visit) we found any of the above

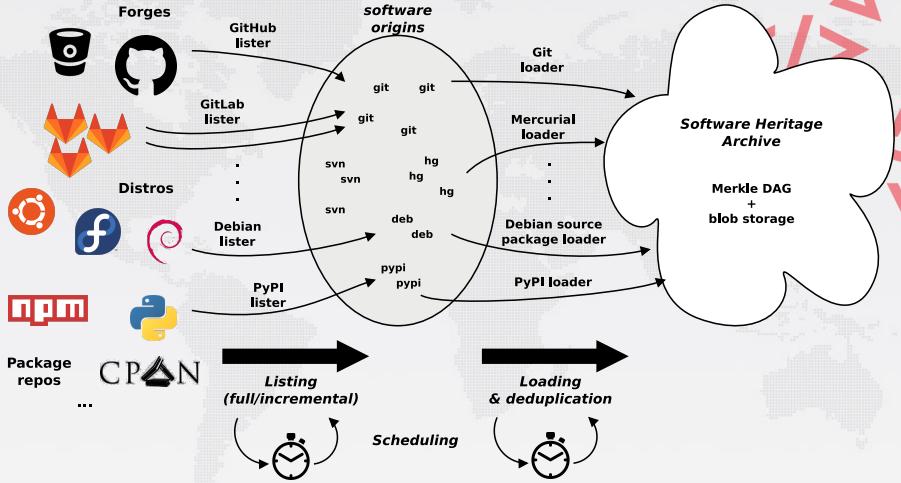
… in a VCS-/archive-agnostic canonical data model

We DON'T archive

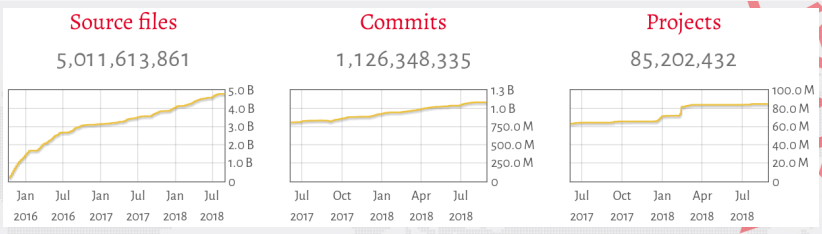
- homepages, wikis
- BTS/issues/code reviews/etc.
- mailing lists

Long term vision: play our part in a *"semantic wikipedia of software"*

Data flow


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

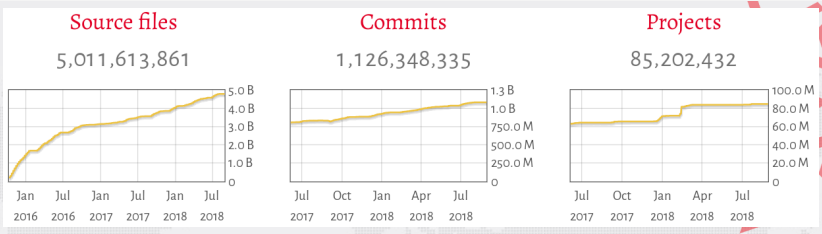
Archive coverage — archive.softwareheritage.org



Current sources

- live: GitHub, Debian, GitLab.com, PyPI
- one-off: Gitorious, Google Code, GNU
- WIP: Bitbucket

Archive coverage — archive.softwareheritage.org

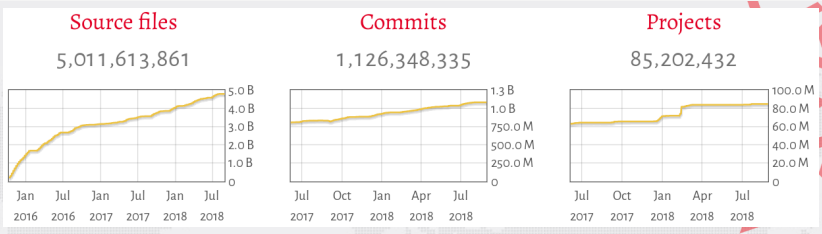


Current sources

- live: GitHub, Debian, GitLab.com, PyPI
- one-off: Gitorious, Google Code, GNU
- WIP: Bitbucket

200 TB (compressed) blobs, 6 TB database (as a graph: 10 B nodes + 100 B edges)

Archive coverage — archive.softwareheritage.org



Current sources

- live: GitHub, Debian, GitLab.com, PyPI
- one-off: Gitorious, Google Code, GNU
- WIP: Bitbucket

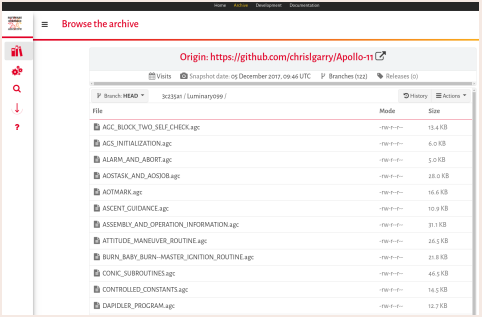
200 TB (compressed) blobs, 6 TB database (as a graph: 10 B nodes + 100 B edges)

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

Demo: the Apollo 11 source code

Margaret Hamilton

The Apollo 11 source code in SWH



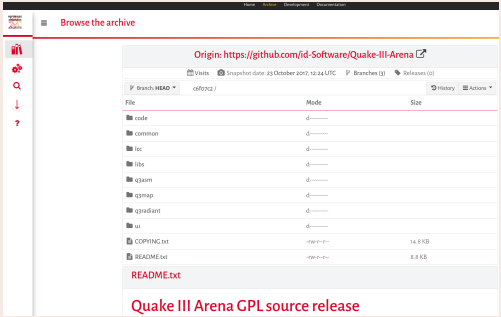
Links

- Entry point
- Burn, baby, burn!

Demo: the Quake 3 source code

John Carmack

The Quake 3 source code in SWH



Links

- Entry point
- What the f…

Web API

RESTful API to programmatically access the Software Heritage archive
`https://archive.softwareheritage.org/api/`

Features

- pointwise browsing of the archive
  - … snapshots ⟶ revisions ⟶ directories ⟶ contents …
- full access to the metadata of archived objects
- crawling information
  - *when have you last visited this Git repository I care about?*
  - *where were its branches/tags pointing to at the time?*

Endpoint index

`https://archive.softwareheritage.org/api/1/`

Bulk download

Vault service

- source code is thoroughly deduplicated within the Software Heritage archive
- bulk download of large artefacts (e.g., a Linux kernel release) requires collecting millions of objects
- the Software Heritage Vault cooks and caches source code bundles for bulk download needs

Tech bits

- RESTful API to request downloads, notifications, and monitoring
- `docs.softwareheritage.org/devel/swh-vault`

Other highlights

Over *10 billions intrinsic* identifiers (IDOs) for scientific reproducibility

See our conceptual framework for DIOs and IDOs `bit.ly/swhpidpaper`

Other highlights

Over *10 billions intrinsic* identifiers (IDOs) for scientific reproducibility

See our conceptual framework for DIOs and IDOs `bit.ly/swhpidpaper`

Research software deposit

- moderated via HAL
  open since September 2018

Other highlights

Over *10 billions intrinsic* identifiers (IDOs) for scientific reproducibility

See our conceptual framework for DIOs and IDOs `bit.ly/swhpidpaper`

Research software deposit

- moderated via HAL

open since September 2018

Compliance deposit

Complete & Corresponding Source code (CCS) deposit for copyleft software shipped in IT products by hardware/software vendors

upcoming

Other highlights

Over *10 billions intrinsic* identifiers (IDOs) for scientific reproducibility

See our conceptual framework for DIOs and IDOs `bit.ly/swhpidpaper`

Research software deposit

- moderated via HAL

open since September 2018

Compliance deposit

Complete & Corresponding Source code (CCS) deposit for copyleft software shipped in IT products by hardware/software vendors

upcoming

Reference archive

See for example

swmath.org

Other highlights

Over *10 billions intrinsic* identifiers (IDOs) for scientific reproducibility

See our conceptual framework for DIOs and IDOs `bit.ly/swhpidpaper`

Research software deposit

- moderated via HAL

open since September 2018

Compliance deposit

Complete & Corresponding Source code (CCS) deposit for copyleft software shipped in IT products by hardware/software vendors

upcoming

Reference archive

See for example

swmath.org

Collaboration hub

- industry, research
- digital preservation

Other highlights

Over *10 billions intrinsic* identifiers (IDOs) for scientific reproducibility

See our conceptual framework for DIOs and IDOs `bit.ly/swhpidpaper`

Research software deposit

- moderated via HAL

open since September 2018

Compliance deposit

Complete & Corresponding Source code (CCS) deposit for copyleft software shipped in IT products by hardware/software vendors

upcoming

Reference archive

See for example

swmath.org

Collaboration hub

- industry, research
- digital preservation

Now part of the French National Plan for Open Science

Reduce risk, avoid fragmentation

Reduce risk, avoid fragmentation

Thomas Jefferson, February 18, 1791

…let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.

Reduce risk, avoid fragmentation

Thomas Jefferson, February 18, 1791

. . . let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.

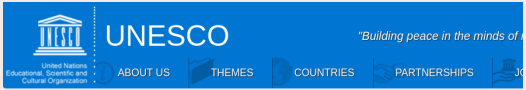
A *common* infrastructure

- mutualisation for sustainability
- open source, non for profit
- mirror network open to all
- may prevent a useless diaspora

Working with UNESCO

Inria Unesco agreement, April 3rd, 2017

Unesco Inria expert group, November 2018



Home › All News › Experts call for greater recognition of software source code as heritage for sustainable development

Experts call for greater recognition of software source code as heritage for sustainable development

16 November 2018

Come in, we're open!

Software Heritage

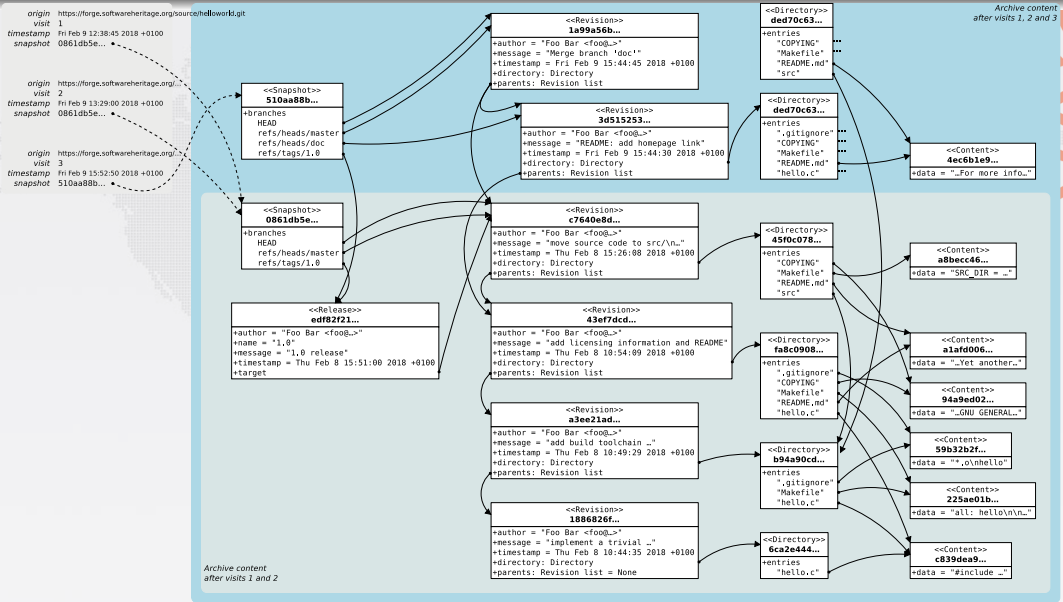
www.softwareheritage.org @swheritage

Everybody is concerned, everybody can help build

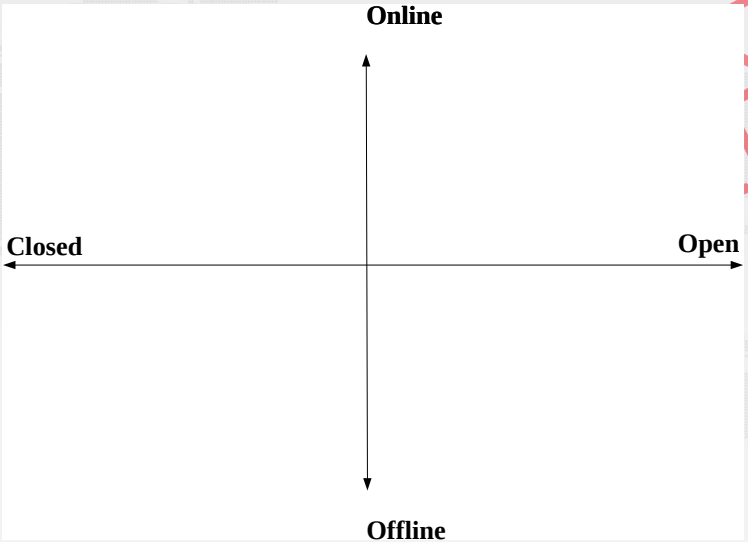
The great library of source code

- preserve the past
- structure the future

The archive: a (giant) Merkle DAG



All the source code



All the source code: strategy