

Software Heritage

building a community to safeguard the Software Commons

Nicolas Dandrimont

Software Heritage

September 10th 2023

DebConf 23



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

- 1 Software and Source Code
- 2 Software Heritage: a mission at the service of Humankind
- 3 Diving deeper into our Features
- 4 Software Heritage Infrastructure
- 5 Concluding remarks



Software is all around us



Software is built from *Source Code*

Harold Abelson, *Structure and Interpretation of Computer Programs* (1st ed.)

1985

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

Software is built from *Source Code*

Harold Abelson, *Structure and Interpretation of Computer Programs* (1st ed.)

1985

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

Apollo 11 source code (excerpt)

```
P63SP0T3      CA      BIT6      # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND   CHAN33
              EXTEND
              BZF    P63SP0T4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF    CODE500      # ASTRONAUT:  PLEASE CRANK THE
              TC     BANKCALL      #              SILLY THING AROUND
              CADR   GOPERF1
              TCF    GOTOP00H      # TERMINATE
              TCF    P63SP0T3      # PROCEED    SEE IF HE'S LYING

P63SP0T4      TC     BANKCALL      # ENTER      INITIALIZE LANDING RADAR
              CADR   SETPOS1

              TC     POSTJUMP      # OFF TO SEE THE WIZARD ...
              CADR   BURNBABY
```

Software is built from *Source Code*

Harold Abelson, *Structure and Interpretation of Computer Programs* (1st ed.)

1985

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

Apollo 11 source code (excerpt)

```
P63SP0T3      CA      BIT6      # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND    CHAN33
              EXTEND
              BZF     P63SP0T4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF     CODE500      # ASTRONAUT:  PLEASE CRANK THE
              TC      BANKCALL     #              SILLY THING AROUND
              CADR    GOPERF1
              TCF     GOTOP00H     # TERMINATE
              TCF     P63SP0T3     # PROCEED    SEE IF HE'S LYING

P63SP0T4      TC      BANKCALL     # ENTER      INITIALIZE LANDING RADAR
              CADR    SETPOS1

              TC      POSTJUMP     # OFF TO SEE THE WIZARD ...
              CADR    BURNBABY
```

Quake III source code (excerpt)

```
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;
}
```

Software is built from *Source Code*

Harold Abelson, *Structure and Interpretation of Computer Programs* (1st ed.)

1985

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

Apollo 11 source code (excerpt)

```
P63SP0T3      CA      BIT6      # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND    CHAN33
              EXTEND
              BZF     P63SP0T4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF     CODE500      # ASTRONAUT:  PLEASE CRANK THE
              TC      BANKCALL     #              SILLY THING AROUND
              CADR    GOPERF1
              TCF     GOTOP00H      # TERMINATE
              TCF     P63SP0T3      # PROCEED    SEE IF HE'S LYING

P63SP0T4      TC      BANKCALL     # ENTER      INITIALIZE LANDING RADAR
              CADR    SETPOS1

              TC      POSTJUMP      # OFF TO SEE THE WIZARD ...
              CADR    BURNBABY
```

Quake III source code (excerpt)

```
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;
}
```

Len Shustek, Computer History Museum

2006

“Source code provides a view into the mind of the designer.”

Software source code as a key asset of Humankind

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

6 November 2018



UNESCO, Inria, Software Heritage invite
40 international experts meet in Paris ...

Software source code as a key asset of Humankind

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

16 November 2018



UNESCO, Inria, Software Heritage invite
40 international experts meet in Paris ...



The call is published on February 2019

Software source code as a key asset of Humankind

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

6 November 2018



UNESCO, Inria, Software Heritage invite
40 international experts meet in Paris ...



The call is published on February 2019

“Recognise software source code as a fundamental enabler in all aspects of human endeavour”

(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

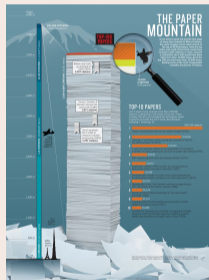
“The real antidote [to the pandemic] is scientific knowledge and global cooperation.”

(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

“The real antidote [to the pandemic] is scientific knowledge and global cooperation.”

Software powers modern scientific research



The top 100 papers

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

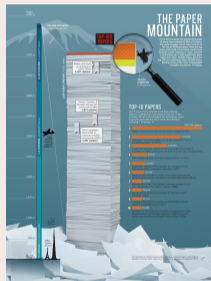
Nature, October 2014

(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

"The real antidote [to the pandemic] is scientific knowledge and global cooperation."

Software powers modern scientific research



The top 100 papers

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

Nature, October 2014

We can still talk to the early inventors



"Telling historical stories is the best way to teach. It's much easier to understand something if you know the threads it is connected to."

Donald E. Knuth
Len Shustek

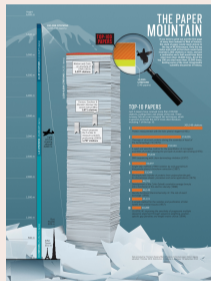
CACM, January 2021

(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

"The real antidote [to the pandemic] is scientific knowledge and global cooperation."

Software powers modern scientific research



The top 100 papers

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

Nature, October 2014

We can still talk to the early inventors



"Telling historical stories is the best way to teach. It's much easier to understand something if you know the threads it is connected to."

Donald E. Knuth
Len Shustek

CACM, January 2021

We need a *dedicated infrastructure* to preserve and share *all* this knowledge!

Enhancing software Reuse, Security and Transparency

Software complexity is growing...

it is important to Know Your SoftWare (KYSW)



Software complexity is growing...

it is important to Know Your SoftWare (KYSW)

Regulation on Software Updates

Recording [...] software versions relevant to a vehicle type

UN Regulations on Cybersecurity, June 2020



Software complexity is growing...

it is important to Know Your SoftWare (KYSW)

Regulation on Software Updates

Recording [...] software versions relevant to a vehicle type

[UN Regulations on Cybersecurity, June 2020](#)

Politique publique de la donnée, des algorithmes et des codes sources

... animer les écosystèmes des... réutilisateurs du source code

[Circulaire du Premier Ministre, 27 Avril 2021, France](#)



Enhancing software Reuse, Security and Transparency

Software complexity is growing...

it is important to Know Your SoftWare (KYSW)

Regulation on Software Updates

Recording [...] software versions relevant to a vehicle type

[UN Regulations on Cybersecurity, June 2020](#)

Politique publique de la donnée, des algorithmes et des codes sources

... animer les écosystèmes des... réutilisateurs du source code

[Circulaire du Premier Ministre, 27 Avril 2021, France](#)

Sec. 4. Enhancing Software Supply Chain Security

ensuring and attesting, to the extent practicable, to the integrity and provenance of open source software

[May 2021 POTUS Executive Order](#)



Enhancing software Reuse, Security and Transparency

Software complexity is growing...

it is important to Know Your SoftWare (KYSW)

Regulation on Software Updates

Recording [...] software versions relevant to a vehicle type

[UN Regulations on Cybersecurity, June 2020](#)

Politique publique de la donnée, des algorithmes et des codes sources

... animer les écosystèmes des... réutilisateurs du source code

[Circulaire du Premier Ministre, 27 Avril 2021, France](#)

Sec. 4. Enhancing Software Supply Chain Security

ensuring and attesting, to the extent practicable, to the integrity and provenance of open source software

[May 2021 POTUS Executive Order](#)



We need a *trusted* knowledge base with *software provenance* !

Software source code is fragile

Endangered source code ...



A word cloud containing the following terms: damage, disaster, malicious, obsolete, attack, aging, media, dependencies, dangling, wear, corruption, encryption, format, deletion, reference, and storage.

- *link rot*: projects are created, moved around, removed
- *data rot*: physical media with legacy software decay
- *platform consolidation* endangers repositories
 - 2015 Google Code and Gitorious.org shutdown: ~1M
 - 2019 Bitbucket mercurial phase out: ~250.000
 - 2022 GitLab.com: **remove inactive projects?**

Software source code is fragile

Endangered source code ...

A word cloud containing the following terms: damage, disaster, malicious, obsolete, attack, aging, media, dependencies, dangling, wear, corruption, encryption, format, deletion, reference, and storage.

- *link rot*: projects are created, moved around, removed
- *data rot*: physical media with legacy software decay
- *platform consolidation* endangers repositories
 - 2015 Google Code and Gitorious.org shutdown: ~1M
 - 2019 Bitbucket mercurial phase out: ~250.000
 - 2022 GitLab.com: **remove inactive projects?**

... is endangered knowledge!

broken links and missing pieces in the *web of knowledge* of humankind

Software source code is fragile

Endangered source code ...



A word cloud containing the following terms: damage, disaster, malicious, obsolete, attack, aging, media, dependencies, dangling, wear, corruption, encryption, format, deletion, reference, and storage.

- *link rot*: projects are created, moved around, removed
- *data rot*: physical media with legacy software decay
- *platform consolidation* endangers repositories
 - 2015 Google Code and Gitorious.org shutdown: ~1M
 - 2019 Bitbucket mercurial phase out: ~250.000
 - 2022 GitLab.com: **remove inactive projects?**

... is endangered knowledge!

broken links and missing pieces in the *web of knowledge* of humankind

Bottomline: we need a global, long term effort

to build a *universal archive* of *all software source code*
and make it *sustainable*

- 1 Software and Source Code
- 2 Software Heritage: a mission at the service of Humankind
- 3 Diving deeper into our Features
- 4 Software Heritage Infrastructure
- 5 Concluding remarks



Unveiled in 2016



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Unveiled in 2016



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog



find and reference all
software source code

Unveiled in 2016



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog



find and reference all software source code

Universal archive



preserve and share all software source code

Unveiled in 2016



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog



find and reference all software source code

Universal archive



preserve and share all software source code

Research infrastructure



enable analysis of all software source code

Today: a *universal* software archive, as a shared infrastructure

One infrastructure
open and shared

Cultural Heritage



Industry



Research



Public Administration



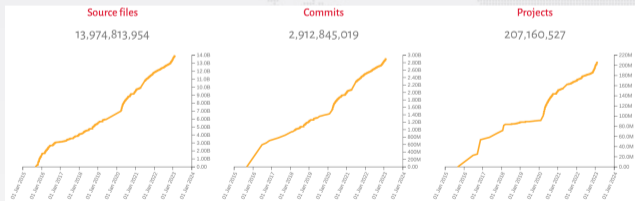
Software Heritage

Today: a *universal* software archive, as a shared infrastructure

One infrastructure
open and shared



The largest archive ever built

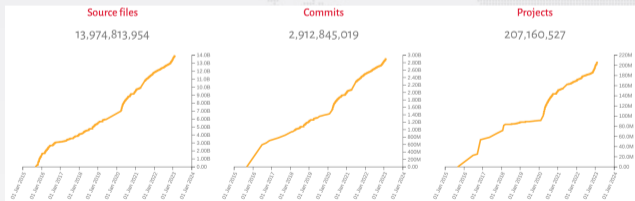


Today: a *universal* software archive, as a shared infrastructure

One infrastructure
open and shared



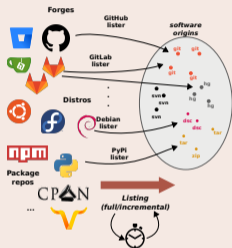
The largest archive ever built



Bitbucket 2,012,133 origins	git 19,494 origins	R 21,486 origins
debian 129,217 origins	gn 6,424 origins	GitHub 152,282,093 origins
GitLab 3,989,638 origins	Guix 12,451 origins	GNU 354 origins
heptapod 1,096 origins	launchpad 356,873 origins	Maven 93,710 origins
NixOS 12,451 origins	npm 1,799,296 origins	Ubuntu 4,083 origins
Phabricator 185 origins	puthon 427,135 origins	SOURCEFORGE 308,970 origins

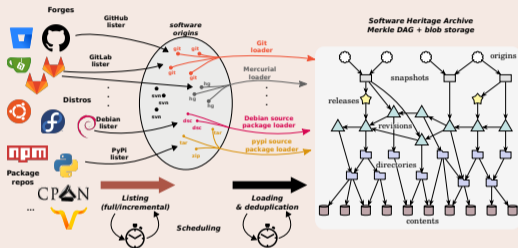
An operational, evolving infrastructure

Harvest and archive



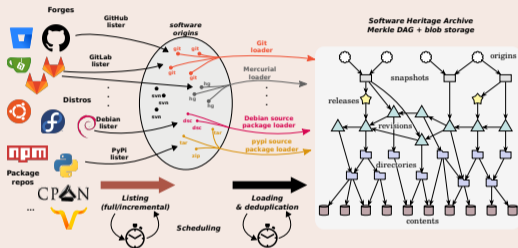
An operational, evolving infrastructure

Harvest and archive



An operational, evolving infrastructure

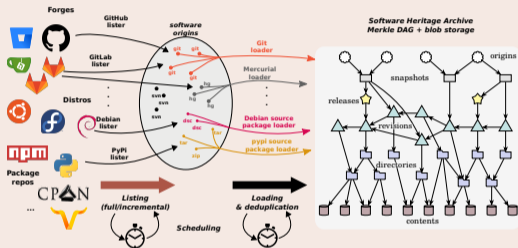
Harvest and archive



- save.softwareheritage.org
- deposit.softwareheritage.org

An operational, evolving infrastructure

Harvest and archive



- save.softwareheritage.org
- deposit.softwareheritage.org

Reference (25 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers



Now in SPDX 2.2, Wikidata, ISO is coming

Growing adoption

Adoption in Open Science

reference archive (example)
for research software

Adoption in Open Science

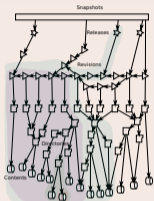
reference archive (example)
for research software

Adoption in Industry and Public Administration

reference archive and knowledge base
for open source software

A revolutionary infrastructure

The *graph* of public software development



All software development
in a **single graph** ...

- enable traceability



A revolutionary infrastructure

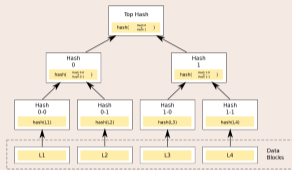
The *graph* of public software development



All software development
in a **single graph** ...

- enable traceability

The *global ledger* of public code



... a **Merkle** graph

- ensure integrity

A revolutionary infrastructure

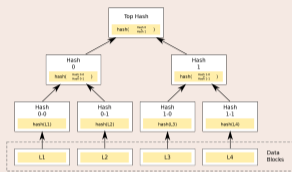
The *graph* of public software development



All software development
in a **single graph** ...

- enable traceability

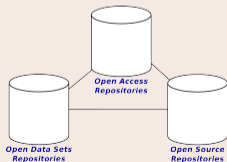
The *global ledger* of public code



... a **Merkle** graph

- ensure integrity

A *pillar* of Open Science



Reference **archive** of
Research Software

- reproducibility
- reference

A revolutionary infrastructure

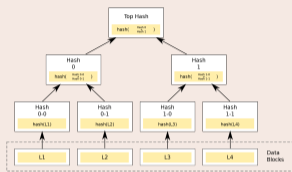
The *graph* of public software development



All software development
in a **single graph** ...

- enable traceability

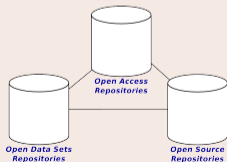
The *global ledger* of public code



... a **Merkle** graph

- ensure integrity

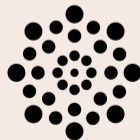
A *pillar* of Open Science



Reference **archive** of
Research Software

- reproducibility
- reference

Reference platform for *Big Code*



uniform data structure

- large scale studies
- machine learning, AI, ...

A walkthrough

General

- Browse [the archive](#), get and use SWHIDs, e.g. [Apollo 11 excerpt](#), [Quake III excerpt](#)
- [Trigger archival](#) in one click with the [browser extension](#)

Open Science

- [Curated deposit via HAL](#), e.g.: [LinBox](#), [SLALOM](#), [Givaro](#), [SumGra](#), [Coq proof](#), ...
- Cite software [with the biblatex-software style](#), e.g.: [article from IPOL](#)

History of software: rescuing landmark legacy software

see [SWHAP process](#), [Software Stories](#), and [SWHAP Days 2022](#)

Public code

Archived source code from [code.gouv.fr](#)

Sharing the vision



United Nations
Educational, Scientific and
Cultural Organization



And many more ...

www.softwareheritage.org/support/testimonials

Sharing the vision



United Nations
Educational, Scientific and
Cultural Organization



And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Inria

Diamond sponsor



Platinum sponsors



Gold sponsors

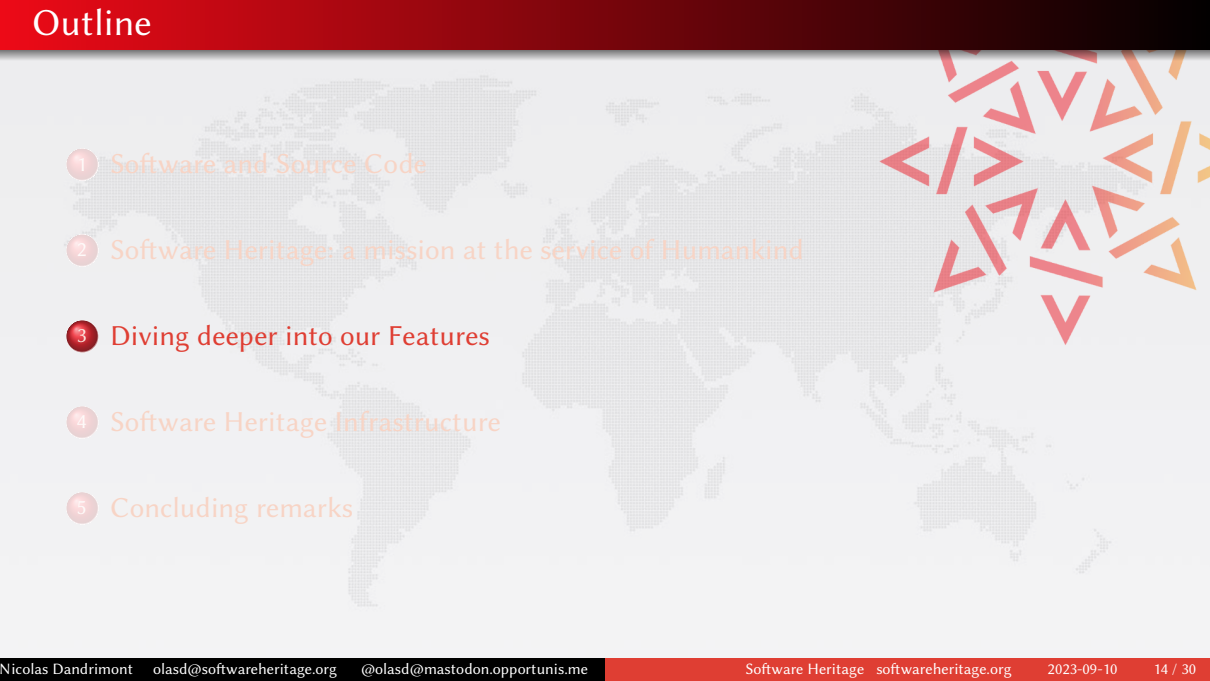


Silver sponsors



Bronze sponsors



- 
- 1 Software and Source Code
 - 2 Software Heritage: a mission at the service of Humankind
 - 3 Diving deeper into our Features
 - 4 Software Heritage Infrastructure
 - 5 Concluding remarks

Object Storage

Challenge

- Storage of and efficient access to the individual versions of archived source code files
- 15 billion objects, median size of 3 kB

Implementation

- multiple backends implementing a **common interface** (fs, Ceph, public clouds)
- **object packing on regular distributed block storage**
- maintenance of 3 copies stored at different locations on different backends

Object Storage

Challenge

- Storage of and efficient access to the individual versions of archived source code files
- 15 billion objects, median size of 3 kB

Implementation

- multiple backends implementing a **common interface** (fs, Ceph, public clouds)
- **object packing on regular distributed block storage**
- maintenance of 3 copies stored at different locations on different backends

Graph storage

Challenge

- storing the global history of software development
- resilient storage & efficient traversals with $>10^{10}$ vertices, $>10^{11}$ edges

Implementation

- store the vertices and edges for resilience in a **"simple" Key-Value store** (backed by Cassandra or PostgreSQL)
- **compressed in-memory snapshots** of the graph for traversals

Multiple challenges to the resilience of the infrastructure

- intentional or unintentional destruction
- legal framework changes
- permanence of the umbrella organization

Building a mirror network

- Avoid single point of (organizational) failure
- Host content under different jurisdictions
- Current deployments:
 - [ENEA \(Italy\)](#)
 - [GRNet for EOSC \(Greece\)](#)

Listers

Listing the contents of hosting platforms to make their contents available for archival

- VCS forges (GitHub, GitLab, Gitea, Forgejo, pagure, ...)
- Distributions (Debian-based, Red Hat-based, language ecosystems like PyPI, Rubygems, ...)
- [Documentation of swh.lister](#)

Listers

Listing the contents of hosting platforms to make their contents available for archival

- VCS forges (GitHub, GitLab, Gitea, Forgejo, pagure, ...)
- Distributions (Debian-based, Red Hat-based, language ecosystems like PyPI, Rubygems, ...)
- [Documentation of swh.lister](#)

Loaders

Converting source code, as published, into our common data model and ingesting the data

- VCS (git, Subversion, CVS, bazaar, mercurial), with full development history
- Software releases (as tarballs or individual files) from other distribution platforms
- [Archive coverage](#)

Save code now

- <https://save.softwareheritage.org/>
- Separate, autoscaling infrastructure
- Support for all available VCSes

Save code now

- <https://save.softwareheritage.org/>
- Separate, autoscaling infrastructure
- Support for all available VCSes

Save Code Now APIs

- public access: [Documentation of API](#)
- raised rate-limits available to partners
- [Browser extension](#)

Push-based architecture

- Based on the interoperable SWORD archival protocol
- Push a set of tarballs, receive a pointer to the SWHID of the archived software
- [Documentation of swh.deposit](#)

Push-based architecture

- Based on the interoperable SWORD archival protocol
- Push a set of tarballs, receive a pointer to the SWHID of the archived software
- [Documentation of swh.deposit](#)

Audience

- Academic partners: automatic ingestion of source code deposited alongside open access research papers
- Industry partners: deposit of the complete corresponding source code for GPLv3 compliance in products

Paris Call on Software Source Code

“[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive”

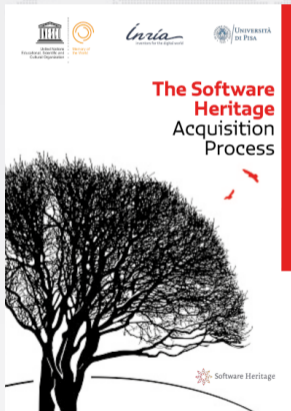
Paris Call on Software Source Code

“[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive”



Paris Call on Software Source Code

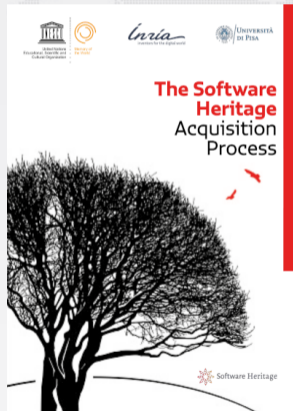
“[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive”



- **Rescue** Legacy Software from different media
 - physical
 - digital
 - legacy / unsupported
 - recent / supported

Paris Call on Software Source Code

“[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive”



- **Rescue** Legacy Software from different media
 - physical
 - digital
 - legacy / unsupported
 - recent / supported
- **Curate** the code
 - reconstructing the development history
 - collecting metadata

Paris Call on Software Source Code

“[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive”



- **Rescue** Legacy Software from different media
 - physical
 - digital
 - legacy / unsupported
 - recent / supported
- **Curate** the code
 - reconstructing the development history
 - collecting metadata
- And **illustrate** with dedicated presentations

Paris Call on Software Source Code

“[We call to] support efforts to gather and preserve the artifacts **and narratives** of the history of computing, while the earlier creators are still alive”

Paris Call on Software Source Code

“[We call to] support efforts to gather and preserve the artifacts **and narratives** of the history of computing, while the earlier creators are still alive”



The screenshot shows the Software Heritage website interface. At the top, it says 'SOFTWARE STORIES' and 'Software Heritage THE GREAT LIBRARY OF SOURCE CODE'. Below that, the main heading is 'The Pisa Collection' with the subtitle 'Stories of landmark legacy code (Beta Version)' and '3 STORIES IN THIS COLLECTION'. There is a search bar labeled 'Search Collection...'. Below the search bar is a photograph of a computer room with a desk, a chair, and a large red cabinet. The caption below the photo is 'TAUmus'.

Paris Call on Software Source Code

“[We call to] support efforts to gather and preserve the artifacts **and narratives** of the history of computing, while the earlier creators are still alive”



The screenshot shows the Software Heritage website interface. At the top, it says 'SOFTWARE STORIES' and 'Software Heritage THE GREAT LIBRARY OF SOURCE CODE'. Below that, the main heading is 'The Pisa Collection' in a large, red font, followed by the subtitle 'Stories of landmark legacy code (Beta Version)' and '3 STORIES IN THIS COLLECTION'. There is a search bar with the text 'Search Collection...'. Below the search bar is a photograph of a computer room with several large, orange-colored cabinets and a desk with a chair. The name 'TAUmus' is written below the photograph.

- **Expand** the SWHAP scope to
 - documents
 - media (videos, pictures, images, etc.)
 - oral history

Paris Call on Software Source Code

“[We call to] support efforts to gather and preserve the artifacts **and narratives** of the history of computing, while the earlier creators are still alive”



The screenshot shows the Software Heritage website interface. At the top, it says 'SOFTWARE STORIES' and 'Software Heritage THE GREAT LIBRARY OF SOURCE CODE'. Below that, the main heading is 'The Pisa Collection' in a large, red font, followed by the subtitle 'Stories of landmark legacy code (Beta Version)' and '3 STORIES IN THIS COLLECTION'. There is a search bar with the text 'Search Collection...'. Below the search bar is a photograph of a computer workstation from the 1970s, featuring a CRT monitor, a keyboard, and a large wooden cabinet. The name 'TAUmus' is written below the photograph.

- **Expand** the SWHAP scope to
 - documents
 - media (videos, pictures, images, etc.)
 - oral history
- **Preserve and Present** all this material

Paris Call on Software Source Code

“[We call to] support efforts to gather and preserve the artifacts **and narratives** of the history of computing, while the earlier creators are still alive”



The screenshot shows the Software Heritage website interface. At the top, it says 'SOFTWARE STORIES' and 'Software Heritage THE GREAT LIBRARY OF SOURCE CODE'. Below that, the main heading is 'The Pisa Collection' with the subtitle 'Stories of landmark legacy code (Beta Version)' and '3 STORIES IN THIS COLLECTION'. There is a search bar labeled 'Search Collection...'. Below the search bar is a photograph of a computer workstation from the 1970s, with a monitor, keyboard, and a large red cabinet. The caption below the photo reads 'TAUmus'.

- **Expand** the SWHAP scope to
 - documents
 - media (videos, pictures, images, etc.)
 - oral history
- **Preserve and Present** all this material
- **Share** process and tools (all open source!)
 - with museums, archives and all interested parties

Paris Call on Software Source Code

“[We call to] support efforts to gather and preserve the artifacts **and narratives** of the history of computing, while the earlier creators are still alive”



- **Expand** the SWHAP scope to
 - documents
 - media (videos, pictures, images, etc.)
 - oral history
- **Preserve and Present** all this material
- **Share** process and tools (all open source!)
 - with museums, archives and all interested parties

see this live on [the Software Stories website](#), and get *the guide*

Discovery of new software origins

Add Forge now

Challenge

discover the wealth of smaller scale hosting platforms based on GitLab, Forgejo, Gitea, Gogs and other self-hostable forge software

Implementation

Submission form on the main archive website for users to submit new forges they've discovered, workflow for ingestion

[Documentation of Add forge now](#)

Add Forge now

Challenge

discover the wealth of smaller scale hosting platforms based on GitLab, Forgejo, Gitea, Gogs and other self-hostable forge software

Implementation

Submission form on the main archive website for users to submit new forges they've discovered, workflow for ingestion

[Documentation of Add forge now](#)

Automation

- dedicated infrastructure for sanity checking the submission, performing the first listing and loading of all the contents
- configurable concurrency (to avoid overloading smaller platforms)
- Work in progress automation of the full process through GitLab pipelines

Graph compression pipeline

Challenge

efficient queries on a graph that has hundreds of billions of edges?

Implementation

- Compress it and hold that representation in memory!
- From tens of terabytes of raw data, to a data structure that can be held in a few hundred GB of RAM
- [Documentation of swh.graph](#)

Graph compression pipeline

Challenge

efficient queries on a graph that has hundreds of billions of edges?

Implementation

- Compress it and hold that representation in memory!
- From tens of terabytes of raw data, to a data structure that can be held in a few hundred GB of RAM
- [Documentation of swh.graph](#)

Dataset exports

- Run your own queries on the full graph of software development
- Available for download, or for direct use on Amazon Athena
- [Documentation of swh.dataset](#)
- [Ethical charter for bulk access](#)

- 1 Software and Source Code
- 2 Software Heritage: a mission at the service of Humankind
- 3 Diving deeper into our Features
- 4 Software Heritage Infrastructure
- 5 Concluding remarks



From humble beginnings



To a very large scale deployment



A very large scale deployment

Our main storage requirements in a nutshell

- More than 1PB of source code files (replicated 3 times by Software Heritage)
- More than 100 TB used for (resilient) storage of the graph
- Infrastructure support for mirroring: 100 TB kafka deployment (~30TB of data used)

A very large scale deployment

Our main storage requirements in a nutshell

- More than 1PB of source code files (replicated 3 times by Software Heritage)
- More than 100 TB used for (resilient) storage of the graph
- Infrastructure support for mirroring: 100 TB kafka deployment (~30TB of data used)

Infrastructure components

- on-site:
 - 4 racks of servers, all running Debian
 - proxmox cluster
 - kubernetes clusters (bare metal and VMs)
 - special purpose clusters (PostgreSQL, kafka, Cassandra, elasticsearch, ...)
 - 1 rack of network equipment

A very large scale deployment

Our main storage requirements in a nutshell

- More than 1PB of source code files (replicated 3 times by Software Heritage)
- More than 100 TB used for (resilient) storage of the graph
- Infrastructure support for mirroring: 100 TB kafka deployment (~30TB of data used)

Infrastructure components

- on-site:
 - 4 racks of servers, all running Debian
 - proxmox cluster
 - kubernetes clusters (bare metal and VMs)
 - special purpose clusters (PostgreSQL, kafka, Cassandra, elasticsearch, ...)
 - 1 rack of network equipment
- off-site:
 - Ceph objstorage cluster: 2 racks
 - Azure resources (blob storage, bare VMs, AKS clusters)
 - AWS resources (S3 for objstorage and dataset exports, Athena for queries)

Low-level deployments

- Everything running Debian, of course
- terraform for provisioning **virtual machines** and **cloud resources**
- puppet for deployment of OS components

Low-level deployments

- Everything running Debian, of course
- terraform for provisioning [virtual machines](#) and [cloud resources](#)
- puppet for deployment of OS components

Kubernetes

- Using [Rancher](#) and [rke2](#) as K8s distribution
- [ArgoCD](#) for continuous deployment ([configuration](#))
- [Helm charts](#) for deploying the [Software Heritage stack](#)
- Jenkins pipelines for image building and Helm chart updates
- Images pushed to the GitLab container registry

A regular Free Software project

- Organized in public
- Browse our GitLab CE instance on gitlab.softwareheritage.org
- Jenkins for CI automation (WIP: GitLab CI)
- real-time discussions on IRC bridged to matrix; open mailing lists
- Browse our documentation on docs.softwareheritage.org

A regular Free Software project

- Organized in public
- Browse our GitLab CE instance on gitlab.softwareheritage.org
- Jenkins for CI automation (WIP: GitLab CI)
- real-time discussions on IRC bridged to matrix; open mailing lists
- Browse our documentation on docs.softwareheritage.org

Infrastructure as code

- All our deployment manifests are published in the open as well
- [Repository access info](#)

A growing and active community

Team

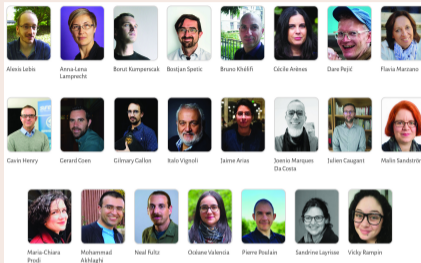


A growing and active community

Team



Ambassadors

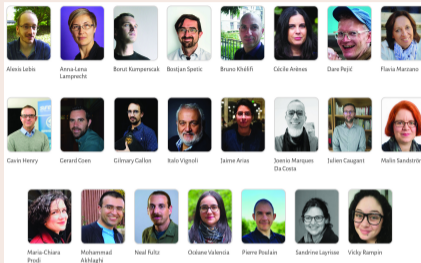


A growing and active community

Team



Ambassadors



Foundations and grantees



- Castalia, CottageLabs
- EasterEggs, OcamlPro
- Octobus, Sperling, Tweag.io
- DataCurrent

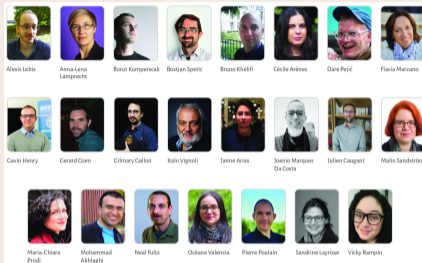


A growing and active community

Team



Ambassadors



Foundations and grantees



- Castalia, CottageLabs
- EasterEggs, OcamlPro
- Octobus, Sperling, Tweag.io
- DataCurrent

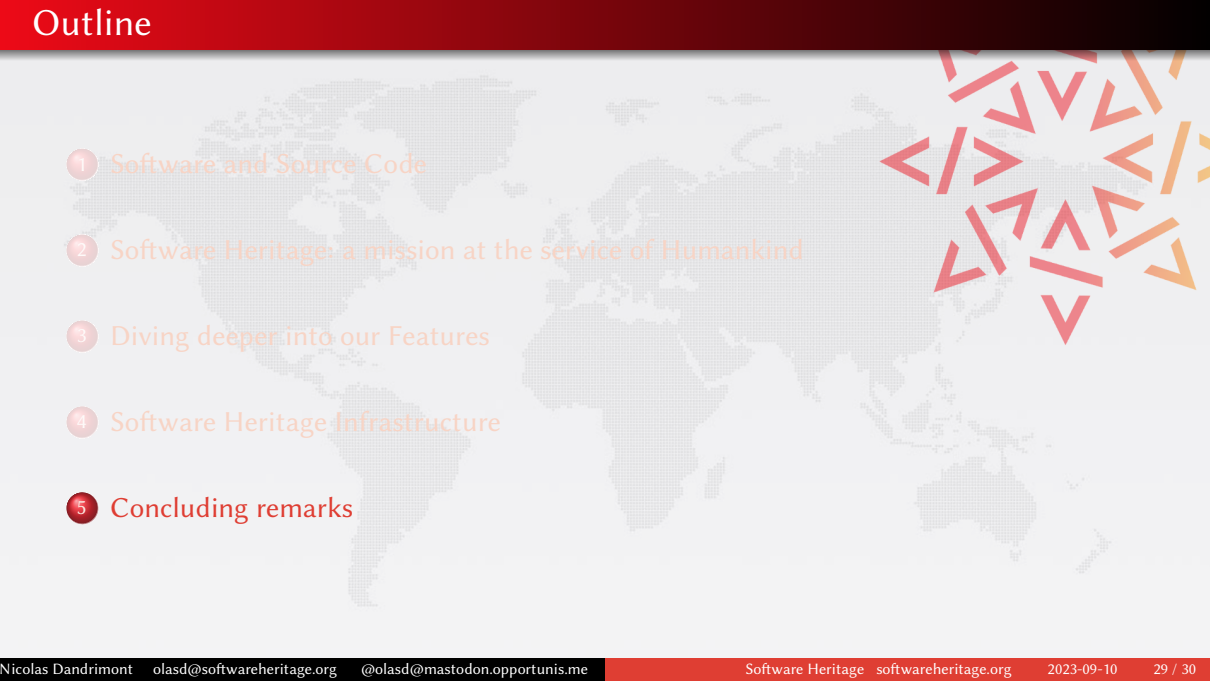
Mirrors and storage partners

“Let us save what remains: ... by such a multiplication of copies, as shall place them beyond the reach of accident.”

— Thomas Jefferson

Enea, GRNET, ...

CEA, RedHat

- 
- 1 Software and Source Code
 - 2 Software Heritage: a mission at the service of Humankind
 - 3 Diving deeper into our Features
 - 4 Software Heritage Infrastructure
 - 5 Concluding remarks

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

Software Heritage is the first brick ...

- **vendor neutral**, multi-stakeholder
- **open source**, **non profit**
- a **worldwide** initiative
- a **long term** initiative

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

Software Heritage is the first brick ...

- **vendor neutral**, multi-stakeholder
- **open source**, **non profit**
- a **worldwide** initiative
- a **long term** initiative

... that will enable

- **archival**, **reference**, **integrity**
- **qualification**, **sharing** and **reuse**
- a **global software knowledge base**
- **test and deploy** **world class tooling**

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

Software Heritage is the first brick ...

- **vendor neutral**, multi-stakeholder
- **open source**, **non profit**
- a **worldwide** initiative
- a **long term** initiative

... that will enable

- **archival**, **reference**, **integrity**
- **qualification**, **sharing** and **reuse**
- a **global software knowledge base**
- test and deploy **world class tooling**

A lot more is needed

Software Heritage can be the *catalyser* of a way bigger undertaking

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

Software Heritage is the first brick ...

- **vendor neutral**, multi-stakeholder
- **open source**, **non profit**
- a **worldwide** initiative
- a **long term** initiative

... that will enable

- **archival**, **reference**, **integrity**
- **qualification**, **sharing** and **reuse**
- a **global software knowledge base**
- test and deploy **world class tooling**

A lot more is needed

Software Heritage can be the *catalyser* of a way bigger undertaking

You can help!

use, adopt, advocate, contribute, fund, support, join

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"



Software Heritage

www.softwareheritage.org

[@swheritage](https://twitter.com/swheritage)

The Library of Alexandria of code



- recover the past
- structure the future
- rebuild trust in science

The Very Large Telescope for Source code



- explore and reuse
- better, more secure software

for society as a whole