

# Archive and Reference Source Code with Software Heritage *a stepping stone for reproducibility*

Roberto Di Cosmo  
Director, Software Heritage  
Inria and Université Paris Cité



Software Heritage  
THE GREAT LIBRARY OF SOURCE CODE

# Software *Source Code* is Precious Knowledge

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)

```
P63SPOT3      CA      BIT6          # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND      CHAN33
              EXTEND
              BZF      P63SPOT4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF      CODE500      # ASTRONAUT:  PLEASE CRANK THE
              TC      BANKCALL      #              SILLY THING AROUND
              CADR      GOPERF1
              TCF      GOTOP00H      # TERMINATE
              TCF      P63SPOT3      # PROCEED      SEE IF HE'S LYING

P63SPOT4      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.”*

# A lightning fast growth

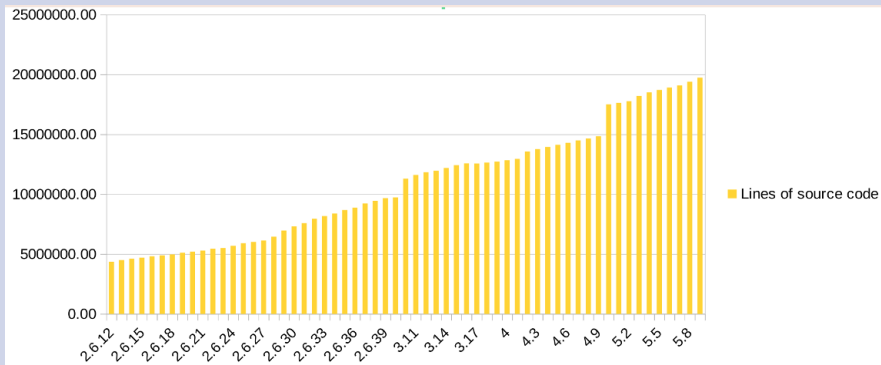
Apollo 11 (~60.000 lines), 1969



"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 : 20+ million lines. . .



. . . now in your pockets!

Open source software is eating the software world

tens of millions of developers collaborate on open source software worldwide today

Reuse is the new rule

80% to 90% of a new application is... just reuse! (Sonatype survey, 2017)

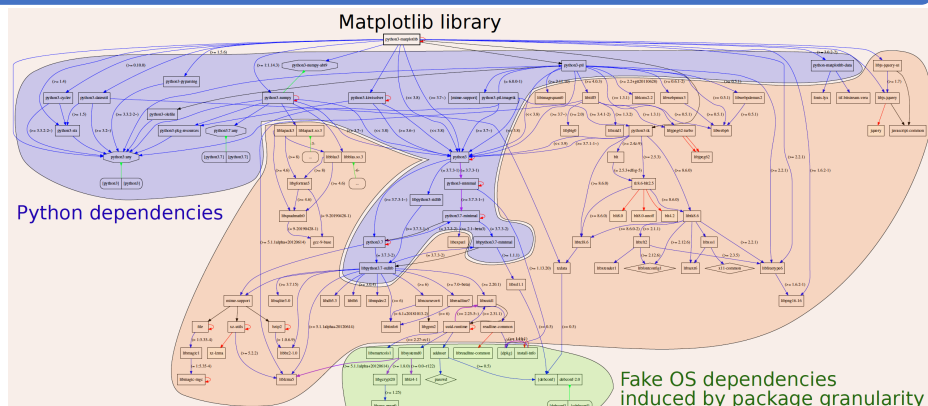
# Source code is *special*: software is *not* data

## Software *evolves* over time

- projects may last decades
- the *development history* is key to its *understanding*

## Complexity

- *millions* of lines of code
- large web of dependencies
  - easy to break, difficult to maintain
  - *research software* a thin top layer
- sophisticated *developer communities*



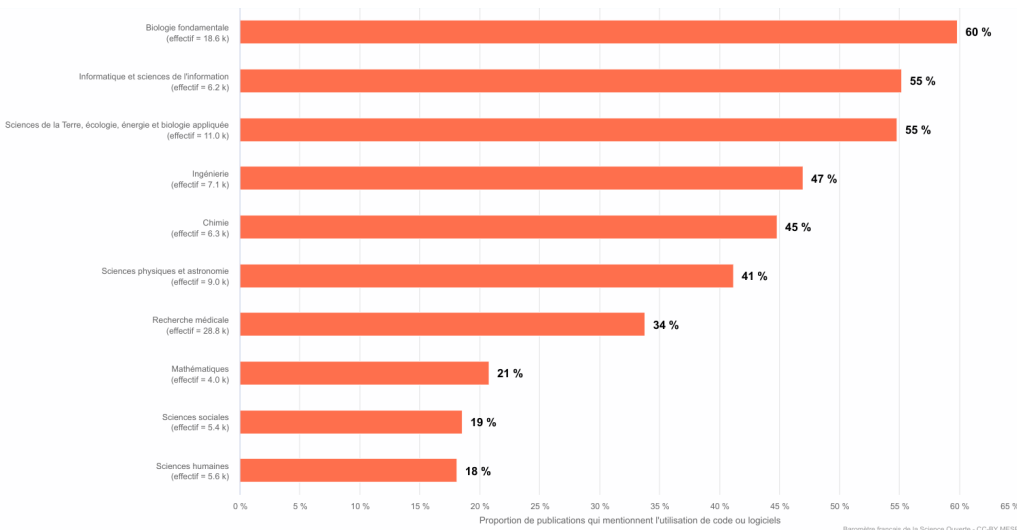
## The human side

design, algorithm, code, test, documentation, community, funding, and so much more...

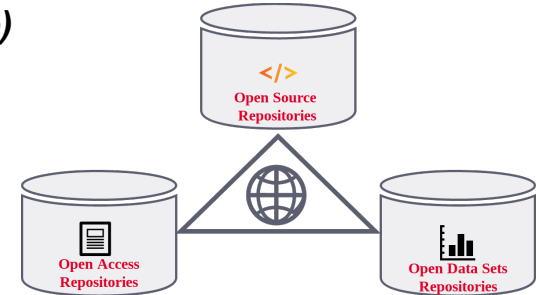
# Software is a pillar of Open Science

## **Software powers all research disciplines!**

Proportion of French publications mentioning use of code or software, by discipline (2024 data from <https://barometredelascienceouverte.esr.gouv.fr/> )



## **A key pillar of Open Science: software (source code)**



← The links in the picture are **important!**

Software may be a **tool**, a **research outcome** and a **research object**

→ Access to the source code is essential!  
→ Preserving (the history of) source code is necessary for **reproducibility**

***How are we handling software and source code in research ?***

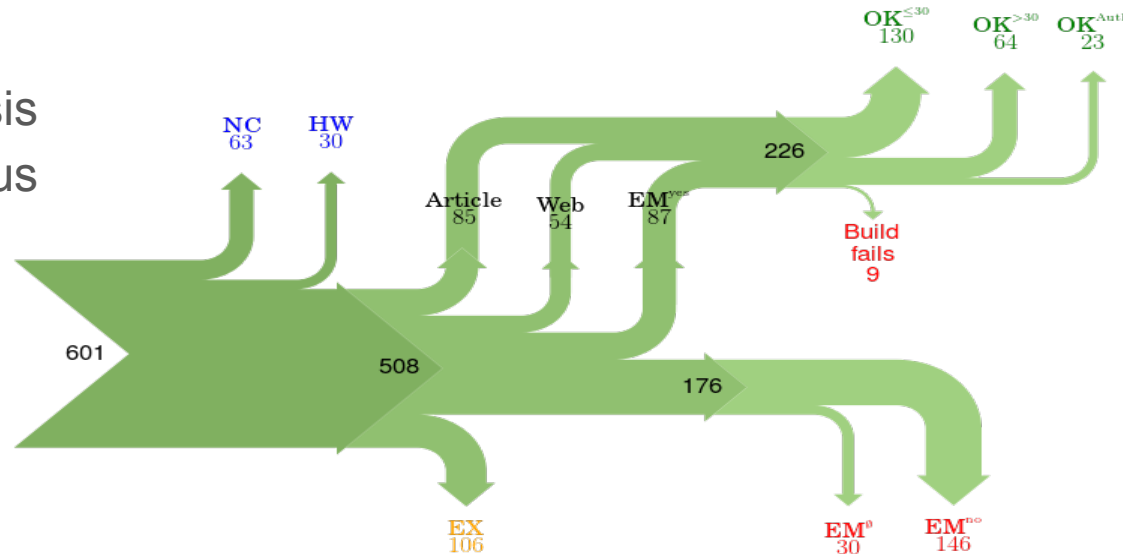
# Reproducibility in Computer Science: state of the art in the field ~2010

## Software Engineering

**2009:** Carlo Ghezzi, 60% of ACM TOSEM papers have code, only 20% installable

## Computer systems research

**2014:** Christian Collberg, analysis of **~600 papers** in prestigious venues, **~200 cannot even find the source code!**



# Awareness and actions

## Artifact Evaluation Committees

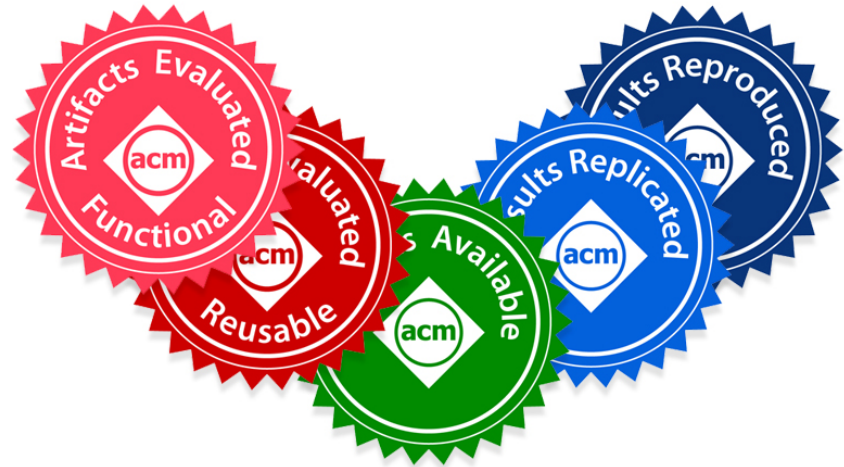
**2011:** run the first time as an award at ESEC-FSE ([J. Vouillon and R. Di Cosmo](#))

**2012-today:** [the process generalizes](#) to a [list too long to maintain](#)

## ACM software badges for publications

See [home page](#) for details.

- Very good intentions, but ...
- Perfectible implementation



A few key issues in reproducibility (*there are many more!*)

## Archive

Ensure **long term availability** of artifacts **with the development history**

## Reference

Ensure **precise identification** of artifacts at **various levels of granularity**

## Describe



Provide **detailed description** (machine readable metadata)

and **proper documentation** (build instructions, dependencies, configuration)

and also *link to relevant papers*



# Not there yet, event for these most basic needs - ACM DL

 Artifacts Available  Artifacts Evaluated & Functional

**Authors/Contributors:** [Authors Info & Affiliations](#)

**DOI:** [https://doi.org/10.1145/\[redacted\]](https://doi.org/10.1145/[redacted]) **version:** 1.0

**Description**

A source archive of [redacted], and the version of [redacted] used in the paper eval. A more up-to-date version of [redacted] can be found at [github.com/\[redacted\]/\[redacted\].git](https://github.com/[redacted]/[redacted].git)

**Assets**

Read Me ([redacted])  
[Download \(3.5 KB\)](#)

Artifact ([redacted])  
[Download \(21.9 MB\)](#)

Only a DOI identifier ...

... does not fill the **author's** needs

Zip file with source code, **looses** the **version control history!**

# Not there yet, event for these basic needs: Papers with code



The Forward-Forward Algorithm: Some Preliminary Investigations

nebulu-ai/nebullvm • PyTorch • NA 2022

The aim of this paper is to introduce a new learning procedure for neural networks and to demonstrate that it works well enough on a few small problems to be worth further investigation.

★ 4,944  
8.64 stars / hour

Paper

Code

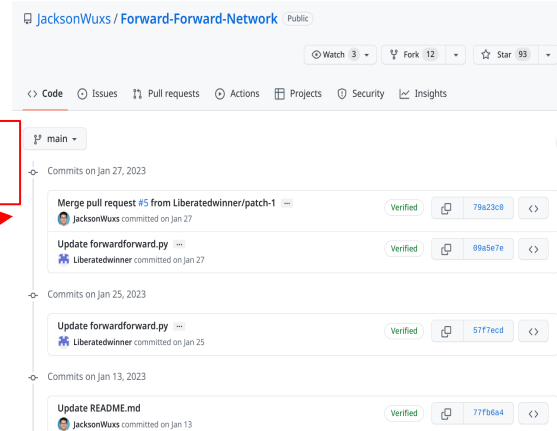
Not an archive!

## Code

Repository	Stars	Framework
nebulu-ai/nebullvm	★ 4,944	PyTorch
keras-team/keras-io	★ 2,128	TensorFlow
mohammadpz/pytorch_forward_forward	★ 1,190	PyTorch
JacksonWuxs/Forward-Forward-Network	★ 93	PyTorch
EscVM/EscVM_YT	★ 48	PyTorch

[See all 6 implementations](#)

Which version?



JacksonWuxs / Forward-Forward-Network Public

main

Commits on Jan 27, 2023

- Merge pull request #5 from Liberatedwinner/patch-1  
Verified 79a23c8
- Update forwardforward.py  
Verified 09a5e7e


Commits on Jan 25, 2023

- Update forwardforward.py  
Verified 5f77ecd

Commits on Jan 13, 2023

- Update README.md  
Verified 77fb6a4


# Forges are not archives!



**SD Times** Latest News Published: March 12th, 2015 - Michael Pehel

## Google begins shutdown of its code repository

After nine years, Google's open-source code repository, Google Code, started closing shop today by disabling new projects and announcing the permanent shut down of the service by January 15, 2015.



**Code collaboration platform GitLab acquires rival Gitorious, will shut it down on June 1**

March 3, 2015 - 4:11 pm

## Sunseting Mercurial support in Bitbucket

April 21, 2020 | 3 min read

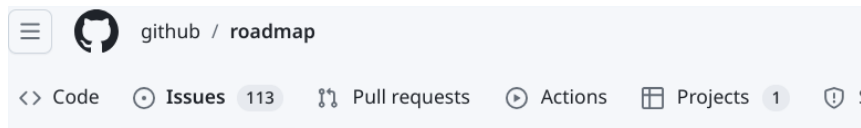
 Denise Chan

Share



[Update Aug 26, 2020] All hg repos have now been disabled and cannot be accessed.

[Update July 1, 2020] Today, mercurial repositories, snippets, and wikis will turn to read-only mode. After July 8th, 2020 they will no longer be accessible.



github / roadmap

<> Code Issues 113 Pull requests Actions Projects 1

### Sunset Subversion support #834

Closed github-product-roadmap opened this issue on Nov 8, 2023 · 1 comment

**Over 1 million projects gone?**

# We need a universal archive Meet Software Heritage!



Cultural Heritage

Industry

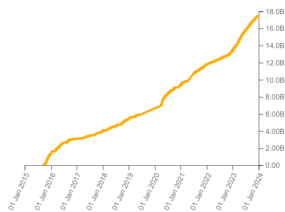
Research

Public Administration



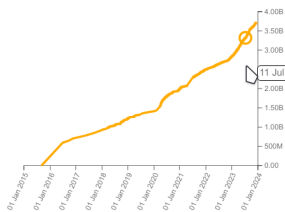
Source files

17,567,724,625



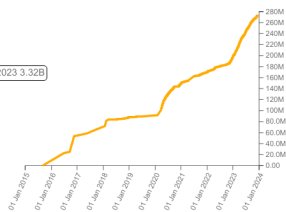
Commits

3,730,352,827



Projects

274,163,348



Directories

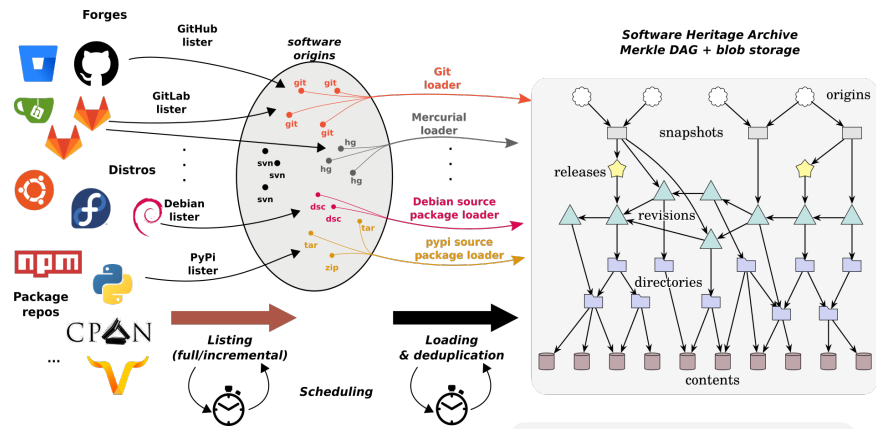
14,101,884,066

Authors

68,895,808

Releases

81,171,066



500+ platforms

All versions history  
in a single graph

- $35 \times 10^9$  nodes
- $500 \times 10^9$  edges
- ~ 1.5 PB storage

Ensures **availability**  
guarantees **integrity**  
enables **traceability**

of all source  
code

**One** common shared **infrastructure**, *replicated*, catering to  
multiple needs



# We can and must do better: archive in Software Heritage



rdicosmo / parmam

Code Issues 4 Pull requests 1 Actions Projects Wiki Security Insights

master

Your master branch isn't protected

Parmam is a minimalistic library allowing to exploit multiverse architecture. Good news: archive is up to date! with its last visit on: 2023-02-28. Click to open the archive page. [rdicosmo.github.io/parmam/](https://rdicosmo.github.io/parmam/)

Roberto Di Cosmo Update biblatex snippet	on Nov 25, 2022	288
config	Use Array.create_float instead of Obj.droppi...	2 years ago
example	Add support for OCaml 5.0	2 months ago
src	Add support for OCaml 5.0	4 months ago
tests	Only run tests/floatscale on 64bit architectures	2 years ago
.gitignore	Version and fix parmam.opam	3 years ago
AUTHORS	Update URL in AUTHORS	4 years ago
CHANGES	Update changelog	2 years ago
LICENSE	Clarified LICENCE and origin of bytearray code.	12 years ago

Releases 12

Update for OCaml 5.0 Latest on Jan 2 +11 releases

Software Heritage Archive

Browse the archive

Enter a SW/HID to resolve or keyword(s) to search for it

<https://github.com/rdicosmo/parmam>

28 February 2023, 01:55:26 UTC

Code Branches (52) Releases (10) Visits

Branch: HEAD 2dc0f46 / History Download Save again

Tip revision: 963608763589e03de38e744d359884d491e65460 authored by Roberto Di Cosmo on 25 November 2022, 20:30:16 UTC

Update biblatex snippet

File	Mode	Size
config		
example		
src		
tests		
.gitignore	-rw-r--r--	38 bytes
AUTHORS	-rw-r--r--	722 bytes

- Regular crawling
- **One click** archival via **updateswh** browser extension
- Webhooks for BitBucket, Gitea, GitHub, GitLab, Sourceforge

 **Gabriel Altay** @gabrielaltay

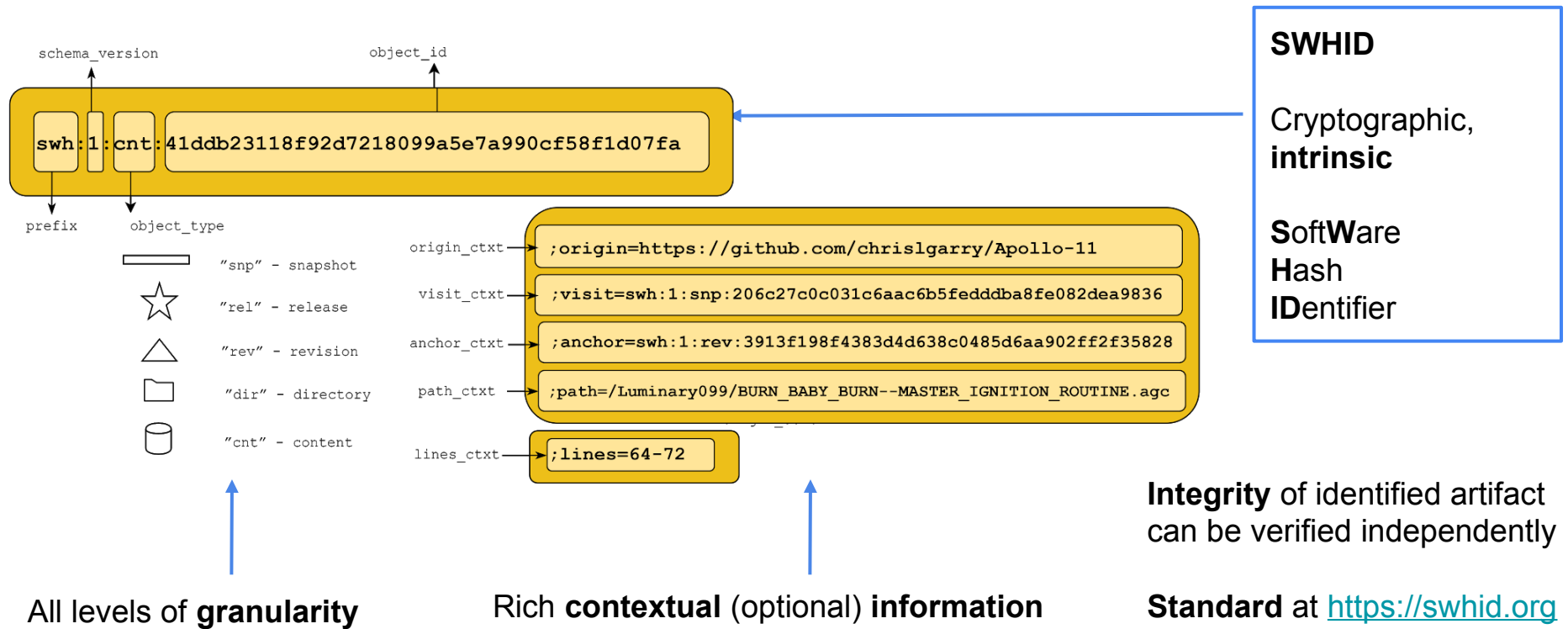
Just realized @Bitbucket disabled all mercurial repositories when the @asclnet informed me that a link associated with an old paper of mine was down. Thought all was lost, but someone archived all the repos! very classy move by @octobus\_net and @SWHeritage.

Traduire le Tweet

1:48 AM · 31 août 2020 · Twitter Web App



# We can and must do better: **reference** in Software Heritage





# We can and must do better: **reference** in Software Heritage

## Getting the SWHID for a code fragment

You can also get the SWHID of a file, or a code fragment inside a file. For this, navigate first to the file, select (optionally) the code fragment of interest by clicking on the line number of the first line, and shift-clicking on the line number of the last line. Then, pull out the red Permalinks tab and copy the SWHID identifier or the corresponding permalink.

~ 30 billion SWHIDs can be found in Software Heritage

```
66
67 let can_redirect path =
68   if not(Sys.file_exists path) then
69     try
70       Unix.mkdir path 0o777; true
71     with Unix.Unix_error(e,_,s) ->
72       (* another job may have created it between the time we checked
73        * if e == Unix.EEXIST then true
74        * else begin
75          (Printf.eprintf "[Pid %d]: Error creating directory '%s'
76           without stdout/stderr\n" (Unix.getpid ()) path (Unix.error_m
77          false
78        * end
79        * else true
80
81
82 let log_debug fmt =
83   Printf.kprintf (
84     if !debug_enabled then begin
85       (fun s -> Format.eprintf "[Parmap]: %s@." s)
86     else ignore
87   ) fmt
88
89 (* freopen emulation, from Xavier's suggestion on OCaml
90 let reopen_out outchan path fname =
91   if can_redirect path then
92     begin
93       flush outchan;
94       let filename = Filename.concat path fname in
95       let fd1 = Unix.descr_of_out_channel outchan in
96       let fd2 = Unix.openfile
```

All levels of **granularity**:

- repository snapshot
- release
- revision
- directory
- file content
- code fragment

# We can do so much better: **reference** in Software Heritage



HOWTO with animations:

<https://www.softwareheritage.org/howto-archive-and-reference-your-code/>

Software Heritage

Mission ▾ Archive ▾ Community ▾ Grants Support us ▾ About ▾ News ▾

## HOWTO archive and reference your code

Archiving and referencing properly your source code is a key principle to comply with the Know Your Software principle (KYSW). This page provides a complete guide to archive and reference your code in Software Heritage.

### Step 1: prepare your public repository

- add a README file
- add an AUTHORS file
- add license information in one of the two recommended ways
  - a LICENSE file at the root of your project, *or*
  - a LICENSES directory containing all the licenses used in your project, and an SPDX compliant copyright header in all your source code files (see the REUSE instructions for details and tools)
- (optionally) add a codemeta.json file containing machine readable metadata (can be produced using the CodeMeta Generator)

Navigation menu items: Features, Browse, Save Code Now, Save Research Software, Save Legacy Code, Browser extensions. Sub-menu items: Benefits, Guidelines (HOWTO).



# A few adoption indicators



## Policy



- [Recommendations in ANR 2023 guidelines \(p. 17\)](#)
- HAL+SWH in [the Open Science software booklet](#)

## Projects



FAIRCORE4EOSC  
Core Components Supporting a FAIR EOSC

The CodeMeta Project



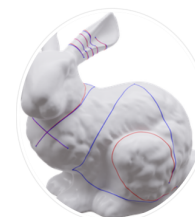
FAIR-IMPACT  
Expanding FAIR solutions across EOSC

## Users and collaborations

### What are they “referencing”?

source	n	percentage
Not available	2868	46.22
GitHub	1151	18.55
software heritage	387	6.24
zenodo	142	2.29
r package	70	1.13
cran	56	0.90
r package version	54	0.87
gitlab	35	0.56

### Graphics Replicability Stamp Initiative

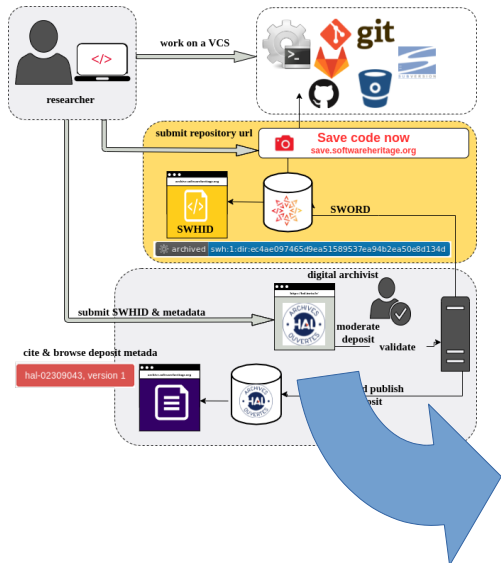


b/Surf: Interactive Bézier Splines on Surface Meshes

Claudio Mancinelli, Giacomo Nazzaro, Fabio Pellacini, Enrico Puppo  
IEEE Transactions on Visualization and Computer Graphics (TVCG)



# In France : HAL + Software Heritage for describe and cite



<https://hal.archives-ouvertes.fr/ha>

**HAL** open science

Free and accessible knowledge

hal-02130801, version 1

LinBox

The LinBox Group 1, 2, 3, 4, 5, 6, 7, 8, 9 Details

- 1 ECO - Exact Computing
- 2 ARIC - Arithmetic and Computing
- 3 AVALON - Algorithms and Software Architectures for Distributed and HPC Platforms
- 4 CIS - Department of Computer and Information Sciences [Newark]
- 5 Drexel University
- 6 NCSU - Department of Mathematics [Raleigh]
- 7 United States Naval Academy
- 8 SCG - Symbolic Computation Group
- 9 CAS3 - Calcul Algébrique et Symbolique, Sécurité, Systèmes Complexes, Codes et Cryptologie

LJK - Laboratoire Jean Kuntzmann

Abstract: LinBox is a C++ template library of routines for solution of linear algebra problems including linear system solution, rank, determinant, minimal polynomial, characteristic polynomial, and Smith normal form. Algorithms are provided for matrices with integer entries or entries in a finite field. A number of matrix storage types is provided, especially for blackbox representation of sparse or structured matrix classes. A few algorithms for rational matrices are available. LinBox also uses underlying data structures and algorithms for integer, rational, polynomial, finite fields and rings, as well as dense and sparse matrix formats coming from the Givaro (<https://caays.gricad-pages.univ-grenoble-alpes.fr/givaro/>) and FFLAS-FFPACK (<http://linbox-team.github.io/fflas-ffpack/>) libraries.

Document type: Software

Domain: Computer Science [cs] Computer Science [cs] / Symbolic Computation [cs.SG]

Complete list of metadata  Display

BROWSE

Software Heritage swh:1:dir:393b611a1424f032e83569b6762502371cfcf65.origin=https://hal.archives-ouvertes.fr/ha-02130801/visit=swh:1:snp:19c29b988fe02623c7076c9b994e02623c7076c9b994e481e691b.anchor=swh:1:rev:e818328952266b7875c692963b11963b1496107.path=1 (hal-02130801)

EXPORT

CodeMeta BibTeX TEI DC DDIterms EndNote

Browse the archive Enter a SWHID to resolve or keyword(s) to search for it

<https://hal.archives-ouvertes.fr/ha-02130801>

14 June 2019, 13:43 UTC

<> Code Branches (1) Releases (0) Visits

Revision: e818328952266b7875c692963b11963b1496107 393b611 / linbox-1.6.3 / linbox / config-blas.h Raw File

Tip revision: e818328952266b7875c692963b11963b1496107 authored by Software Heritage on 11 June 2019, 08:12 UTC hal: Deposit 297 in collection hal

config-blas.h

```

1 /* config-blas.h
2  * Copyright (C) 2005 Pascal Giorgi
3  *          2007 Clement Pernet
4  * Written by Pascal Giorgi <pgiorgi@waterloo.ca>
5  *
6  * =====LICENCE=====
7  * This file is part of the library LinBox.
8  *
9  * LinBox is free software: you can redistribute it and/or modify
10 * it under the terms of the GNU Lesser General Public
11 * License as published by the Free Software Foundation; either
12 * version 2.1 of the License, or (at your option) any later version.
13 *
14 * This library is distributed in the hope that it will be useful,
15 * but WITHOUT ANY WARRANTY; without even the implied warranty of
16 * MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU
17 * Lesser General Public License for more details.
18 *
19 * You should have received a copy of the GNU Lesser General Public
20 * License along with this library; if not, write to the Free Software
21 * Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301 USA
22 * =====LICENCE=====
23
24
25
26 #ifndef LINBOX_config_blas_h

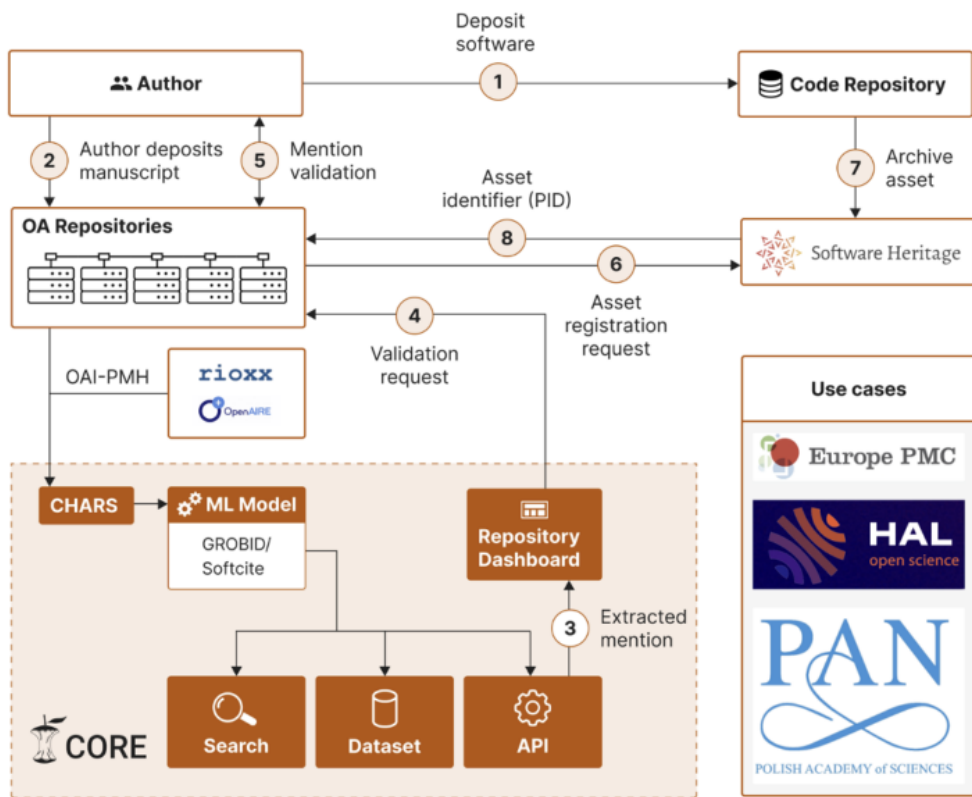
```

[swh:1:dir:393b611a1424f032e83569b6762502371cfcf65](https://hal.archives-ouvertes.fr/ha-02130801/visit=swh:1:snp:19c29b988fe02623c7076c9b994e02623c7076c9b994e481e691b.anchor=swh:1:rev:e818328952266b7875c692963b11963b1496107.path=1)

# Demo time

- [Browse](#) and [Reference](#) (e.g. [Apollo 11 \[excerpt\]](#), your work [may be already there](#) !)
- [Trigger archival](#), use [the updateswh browser extension](#), configure [the webhooks](#)
- Cite with [biblatex-software](#) (CTAN, [Overleaf ACMART template](#))
- Describe with Codemeta (use [codemeta generator](#))
- [Curated deposit in SWH via HAL](#), see for example: [LinBox](#), [SLALOM](#), [Givaro](#), [NS2DDV](#), [SumGra](#), [Coq proof](#), ...
- Extracting all the software products [for Inria](#), [for CNRS](#), [for CNES](#), [for LIRMM](#) or [for Rémi Gribonval](#) using [HalTools](#)
- Example with Parmap: [devel on Github](#), [archive in SWH](#), [curated deposit in HAL](#)
- Example research articles:
  - compare Fig. 1 and conclusions in [the 2012 version](#) and [the updated version](#)
  - SWHID in [a replication experiment](#)

# Latest news : SOFair and SCOSS



## SOFTWARE HERITAGE

### THE LIBRARY OF ALEXANDRIA OF SOFTWARE SOURCE CODE

Software Heritage is an open non-profit infrastructure for archiving, referencing and sharing software source code, launched by Inria in 2016, in partnership with UNESCO.

Archiving over 260 million software projects already, it is built according to the UNESCO recommendations for Open Science: open, multi-stakeholder, non-profit, using exclusively open source components, it serves as a cornerstone for Open Science.

It simplifies the deposit of research software and associated metadata, amplifying the visibility and impact of scholarly outputs. Researchers take advantage of Software Heritage's vast collection of software projects, that enables citation, referencing and sharing of software artefacts, improving reproducibility and traceability of research. Libraries benefit from Software Heritage's robust infrastructure, which offers long-term archival and unique identification of software, removing the need for custom and in-house archival solutions.

By supporting Software Heritage, you're supporting unfettered access, reference and citation of software produced by academic research, reinforcing the principles of open science.

## WHY HAS IT BEEN DEEMED AN ESSENTIAL INFRASTRUCTURE?

The SCOSS Board considers Software Heritage to be an essential open science infrastructure because it provides continued access to the software and code outputs produced by researchers globally.

## SCOSS FUNDING TARGET

€ 900,000

# The way ahead

## Archival and reference for source code

- **Technical barriers** are mostly solved issues (*over 6 years of work*)
- **Social barriers** still stand in the way (adoption, training, cost mutualization, ...)

## Thank you

- Software Heritage: <https://softwareheritage.org> and [the 2022 annual report](#)
- HOWTO archive, reference, describe and cite research software: <https://bit.ly/swh-howto-research>
- Software deposit and metadata curation: [HAL-SWH Webinar, July 2022](#)
- Deuxième plan national pour la Science Ouverte: [official website](#)
- Software Pillar session in OSEC 2022: [official website](#)
- EOSC SIRS report: <https://data.europa.eu/doi/10.2777/28598>
- Roberto Di Cosmo and Marco Danelutto. [Rp] Reproducing and replicating the OCamlP3I experiment. ReScience C, 6(1):#2, April 2020. [link]

Learn more