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     GOTOPOOH        # 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."

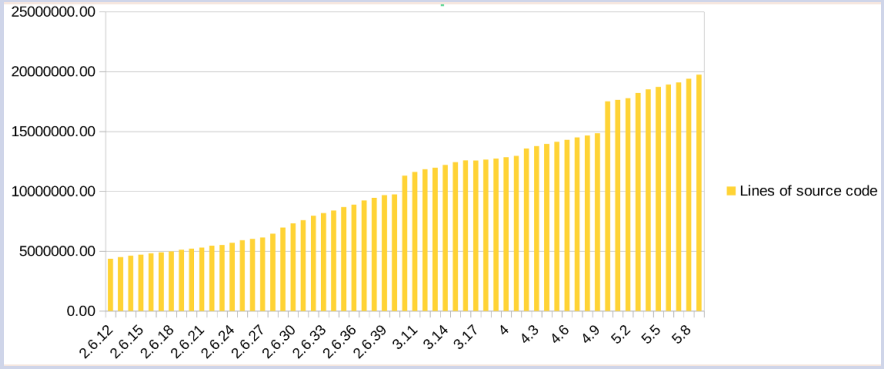
A lightning fast growth

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)

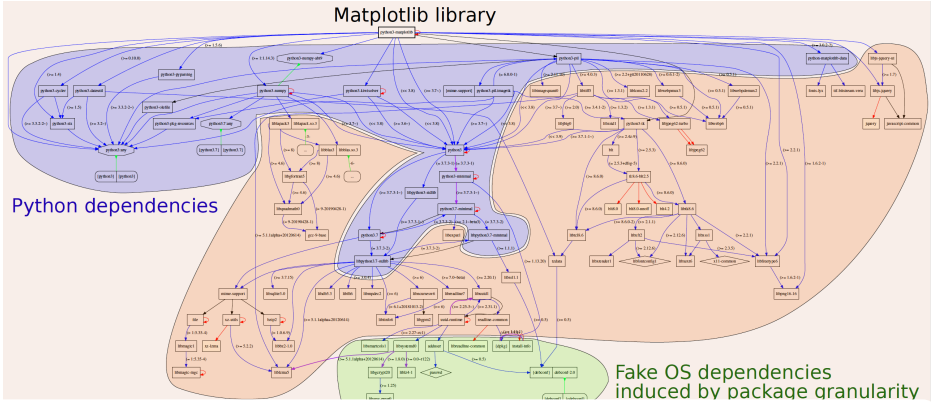
Source code is *special*: software is *not* data

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*

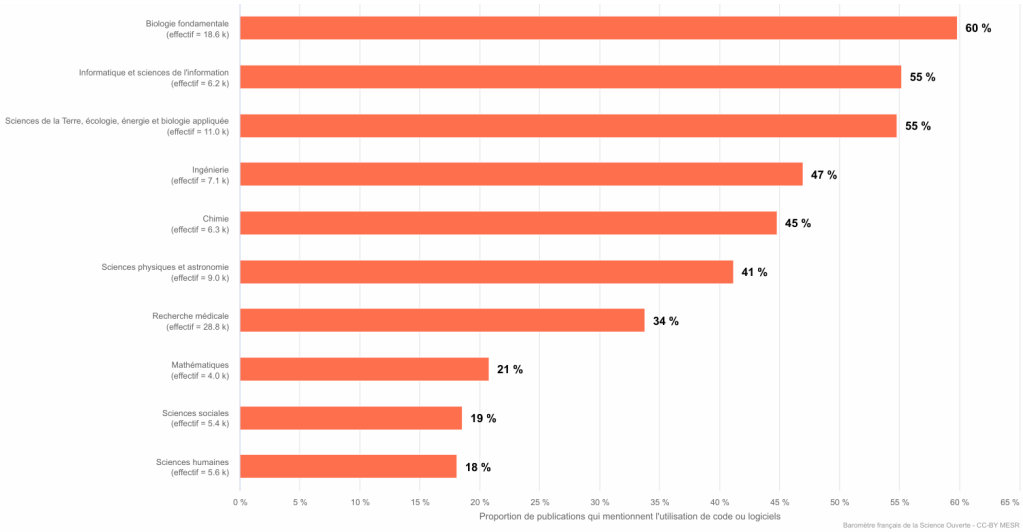


The human side

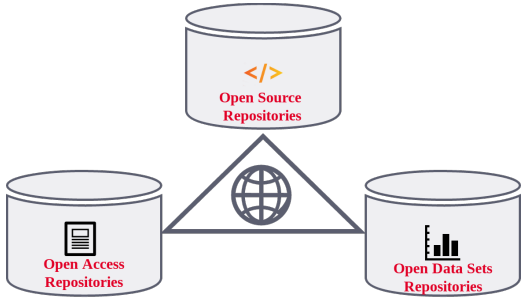
design, algorithm, code, test, documentation, community, funding, and so much more...

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 ?

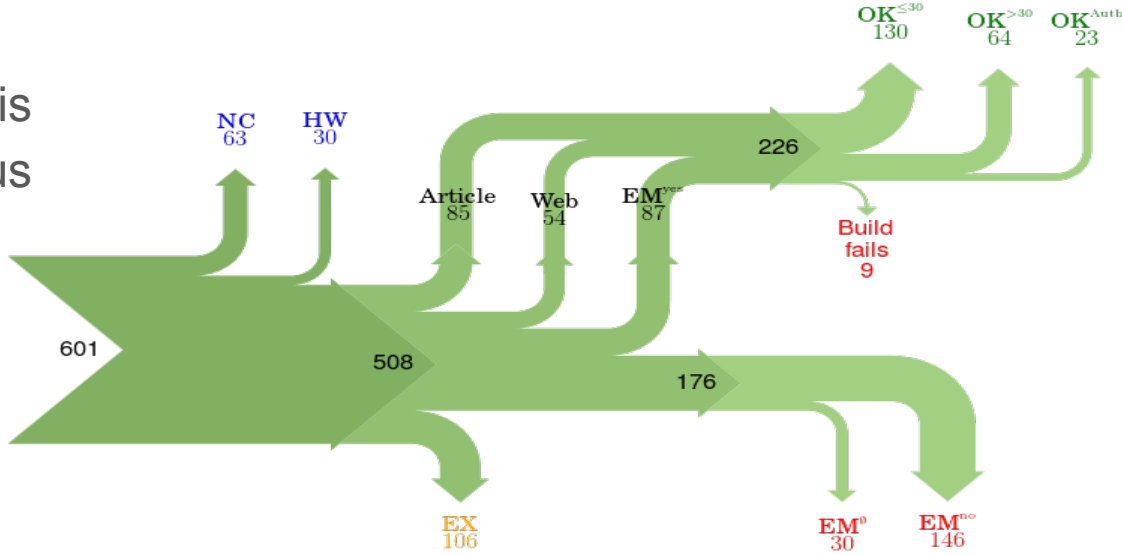
Reproducibility in Computer Science: state of the art in the field ~2010

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

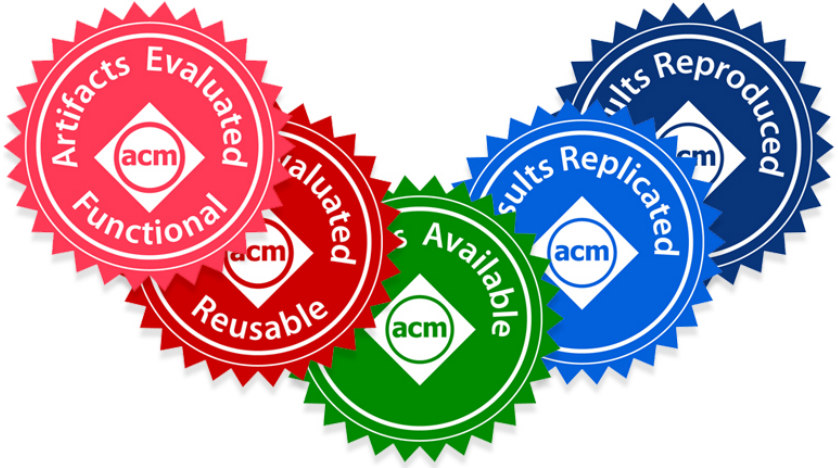
2011: run the first time as an award at ESEC-FSE (J. Vouillon and R. Di Cosmo)

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**

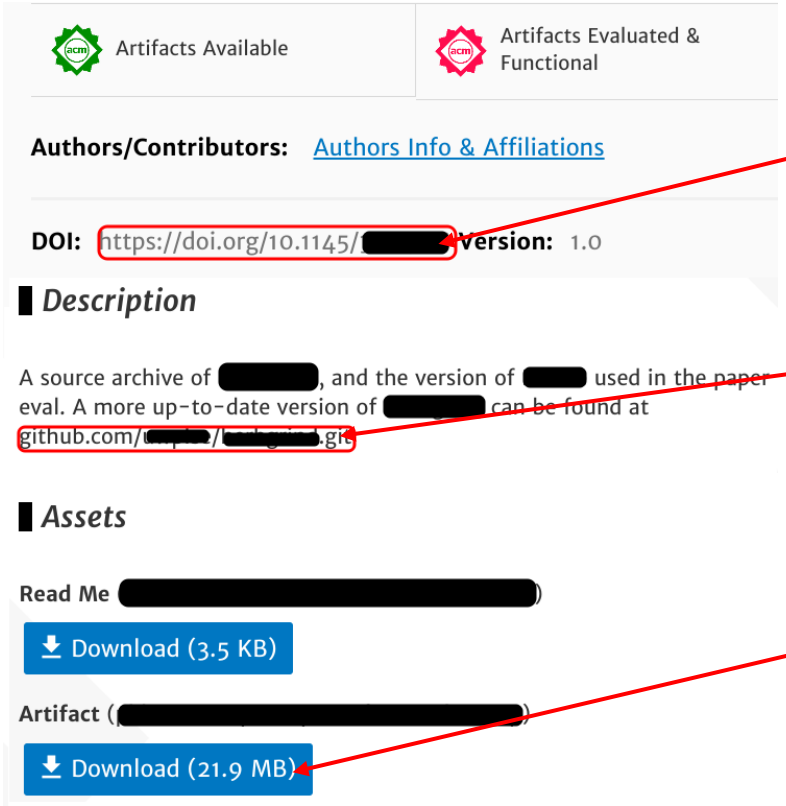
Describe

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



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

nebuly-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!

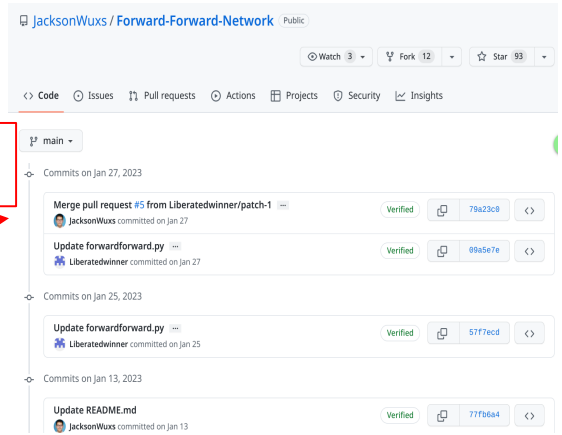
Code

Edit

nebuly-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?



Forges are not archives!

SD Times

Google begins shutdown of its code repository

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

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 Janua

Googl
as Go
Sourc
broug
was c
servic
result

TNW News

Latest Deep tech Sustainability Ecosystems Data and security Fintech and ecommerce Future of work More

This article was published on March 3, 2015

INSIDER

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

March 3, 2015 - 4:11 pm

Bitbucket

Sunsetting 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!

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

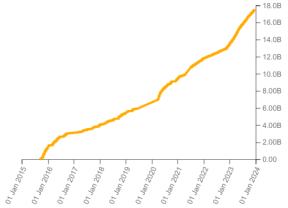
Inria and unesco

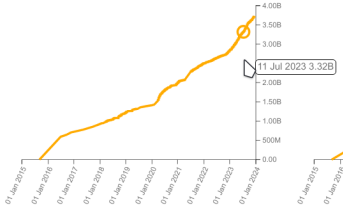
Cultural Heritage | **Industry** | **Research** | **Public Administration**

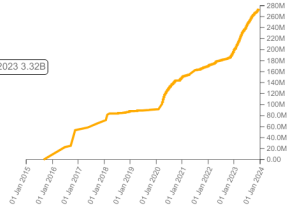
Source files
17,567,724,625

Commits
3,730,352,827

Projects
274,163,348



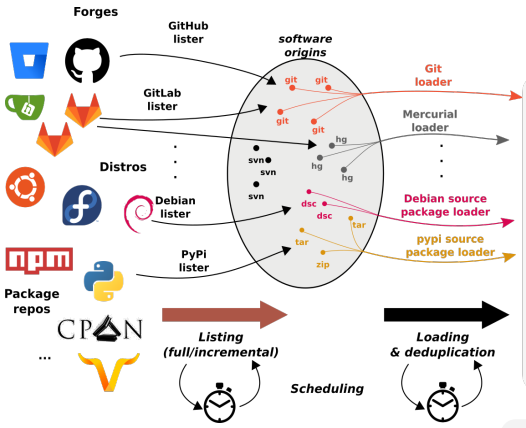




Directories
14,101,884,066

Authors
68,895,808

Releases
81,171,066



500+ platforms

All versions history in a single graph

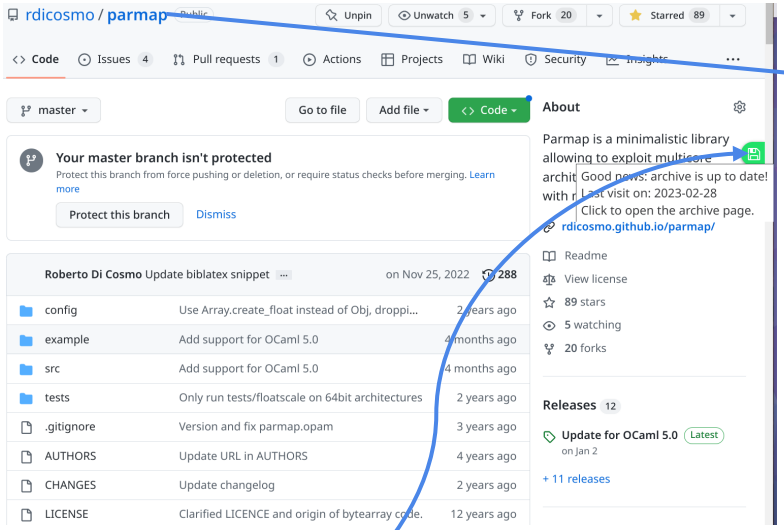
- **35 × 10⁹** nodes
- **500 × 10⁹** edges
- ~ **1.5 PB** storage

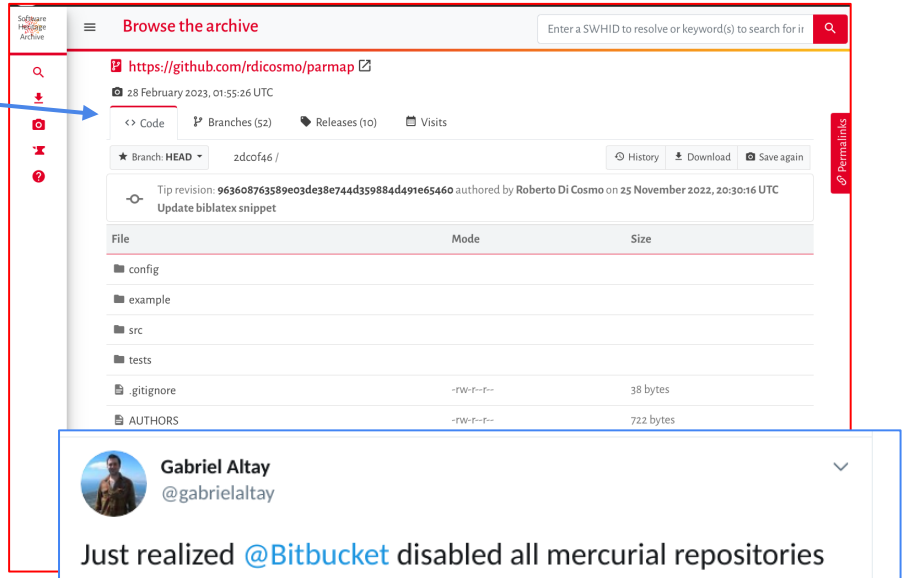
Ensures **availability**
guarantees **integrity**
enables **traceability**

of all source code

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

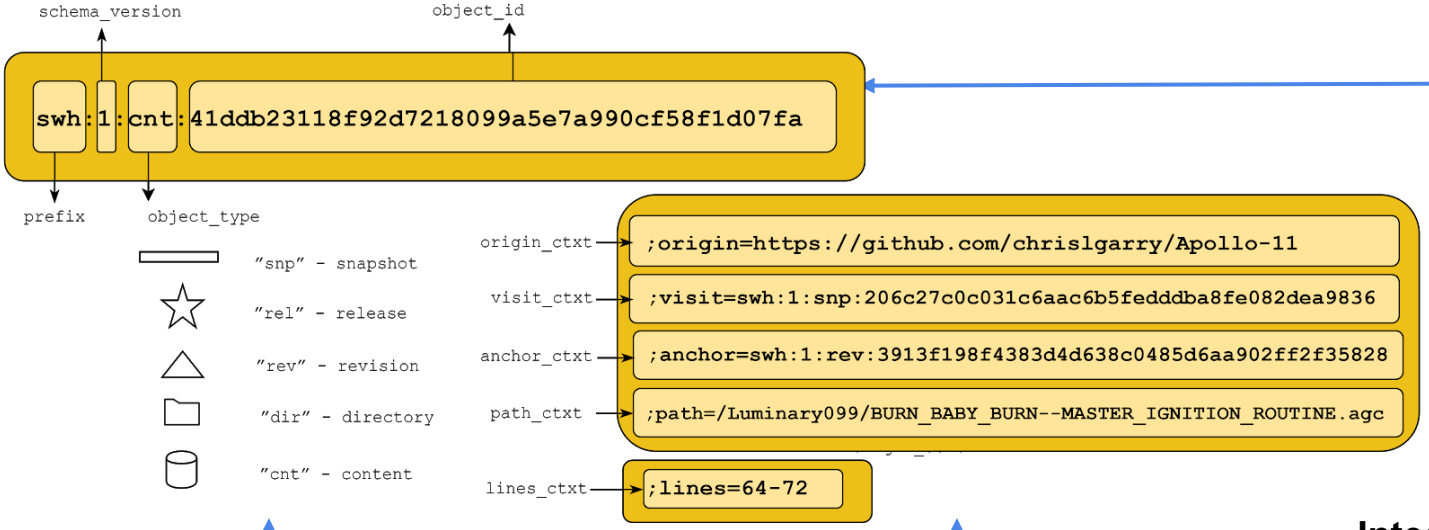
We can and must do better: **archive** in Software Heritage





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

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



SWHID

Cryptographic, **intrinsic**

Soft**W**are
Hash
IDentifier

All levels of **granularity**

Rich **contextual** (optional) **information**

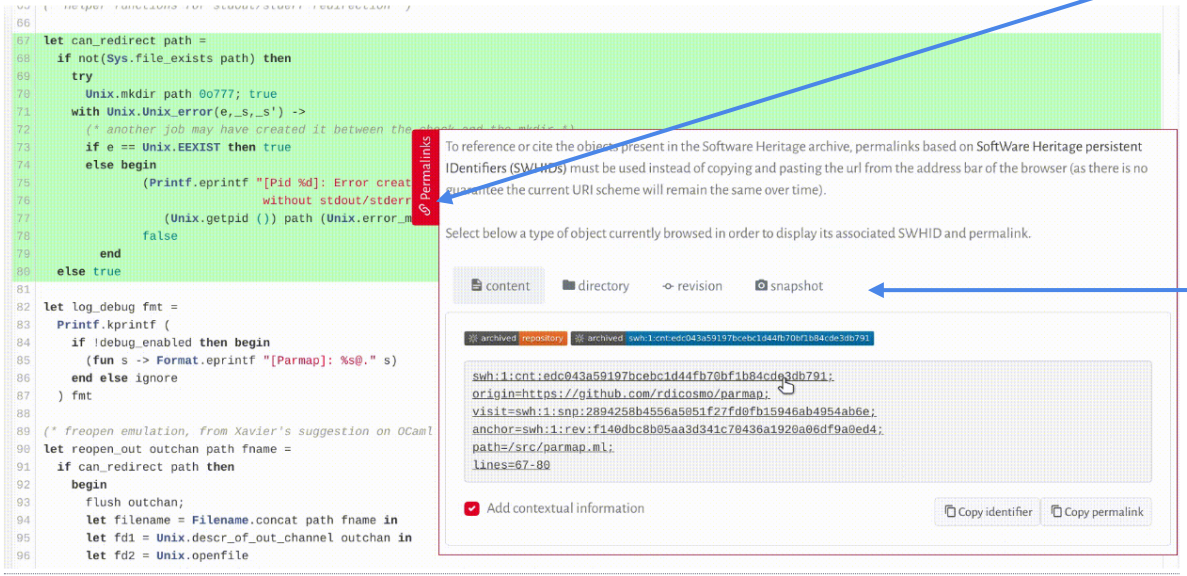
Integrity of identified artifact can be verified independently

Standard at https://swhid.org

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.

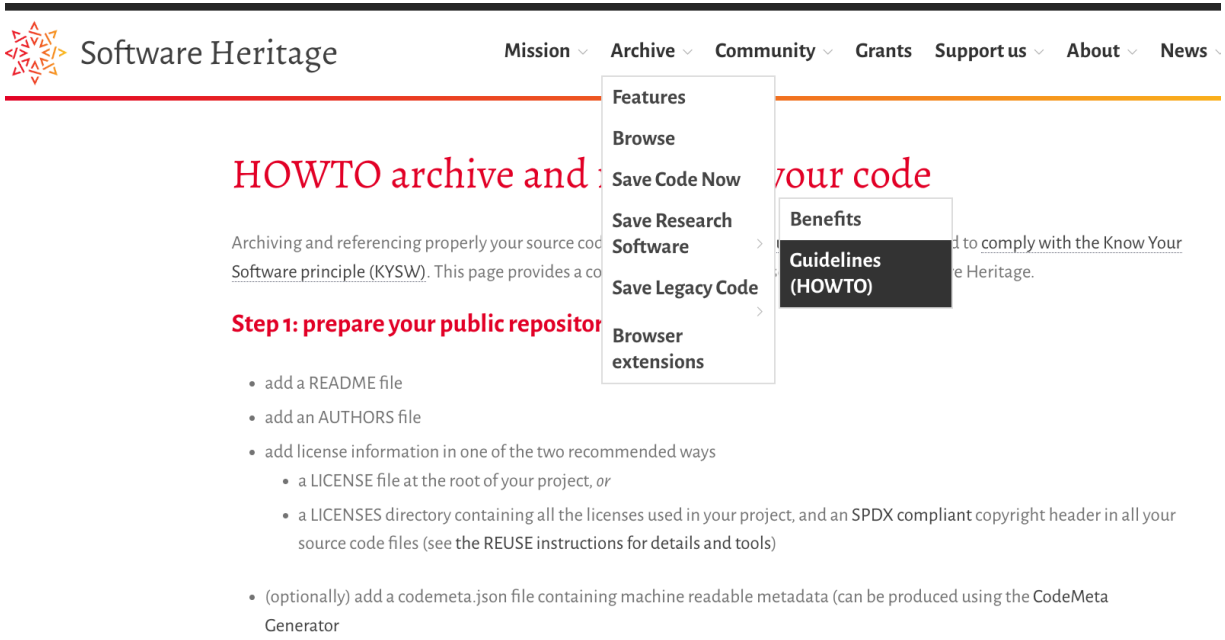


~ 30 billion SWHIDs can be found in Software Heritage

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/



A few adoption indicators

Policy

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

Projects

eosc | FAIRCORE4EOSC
Core Components Supporting a FAIR EOSC

The CodeMeta Project

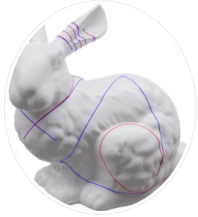
eosc | FAIR-IMPACT
Expanding FAIR solutions across EOSC

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

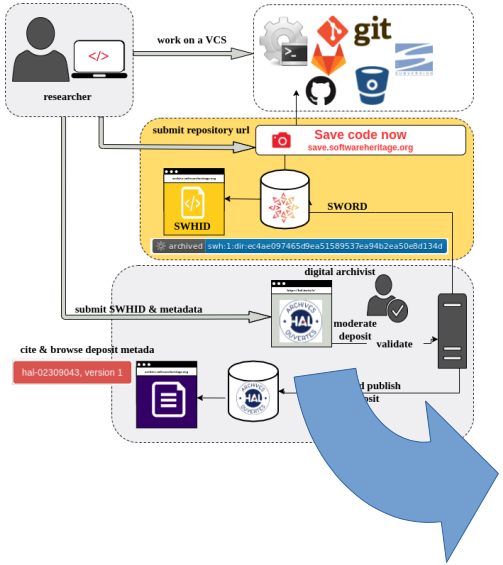


b/Surf: Interactive Bézier Splines on Surface Meshes

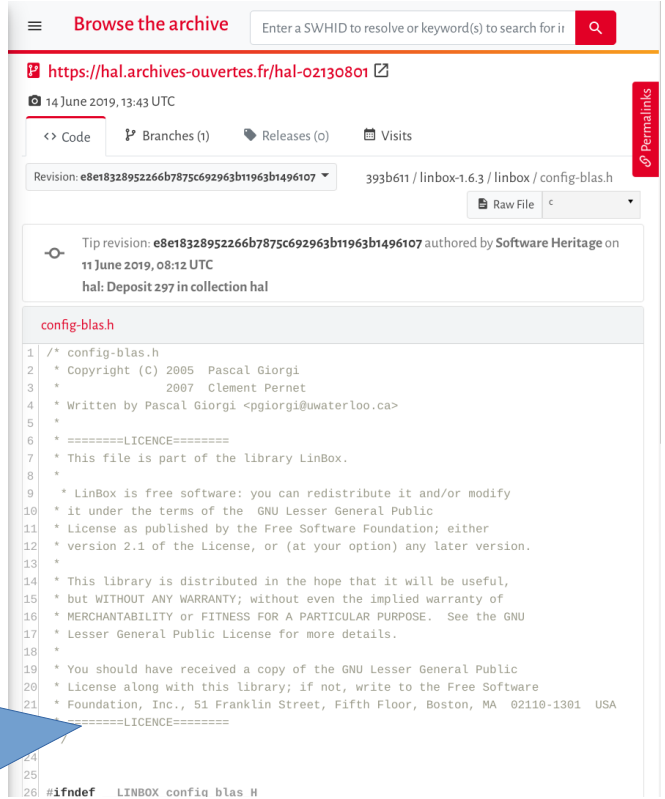
Claudio Mancinelli, Giacomo Nazzaro, Fabio Pellacini, Enrico Puppo
IEEE Transactions on Visualization and Computer Graphics (TVCG)

doi | Repository

In France : HAL + Software Heritage for describe and cite



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

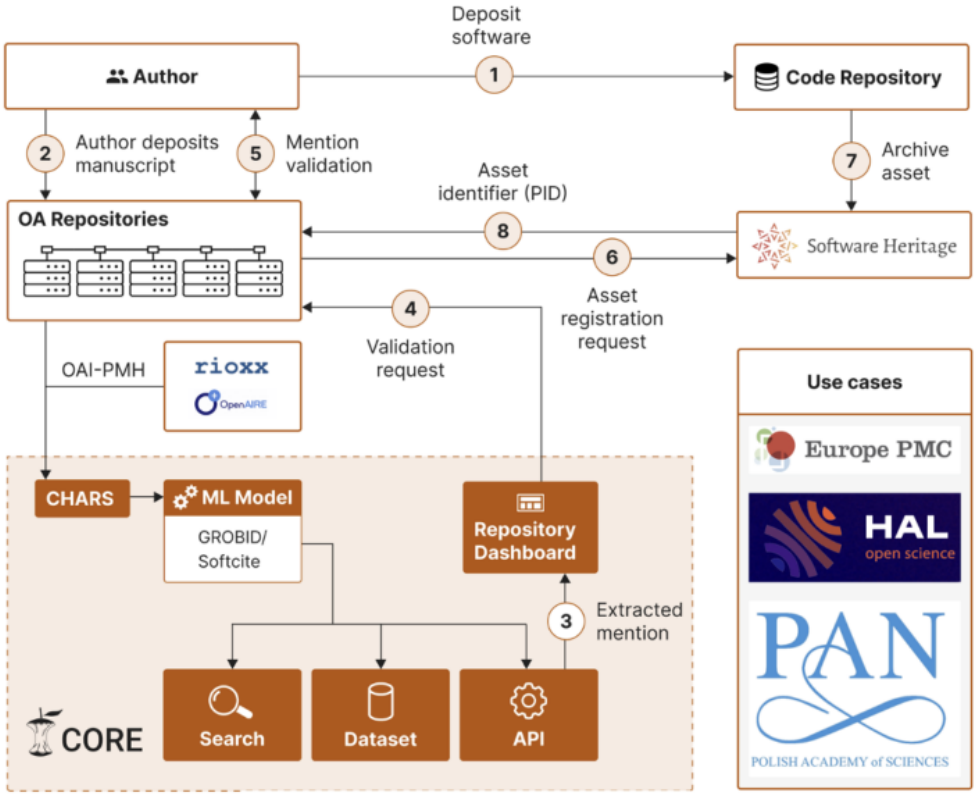


swh:1:dir:393b611a1424f032e83569bf6762
371cfcf65

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



SCOSS

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

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

Learn more

- 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 OCamlP3l experiment. ReScience C, 6(1):#2, April 2020. [link]