The Software (Source Code) pillar of (Open) Science

Software Heritage at your service

Roberto Di Cosmo

Director, Software Heritage
Inria & Université Paris Cité

22 October 2026
Circle U Open Science task force
Université Paris Cité



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Software (Source Code) as a pillar of (Open) Science
2. Software Heritage at your service
3. European research software through the Software Heritage Very Large Telescope
4. What you can take away this morning

Last year, in one slide

22 October 2025, "Software as a Pillar of Open Science"

- source code is precious knowledge, and a pillar of open science
- the archive: what it holds, and how it is built
- *save now*: archive any repository, from any forge
- the SWHID, now ISO/IEC 18670:2025
- citing software: *biblatex-software*
- describing software: CodeMeta
- depositing software in HAL, through Software Heritage

We do a brief recap the rest is one scan/click away.

The slides



annex.softwareheritage.org/public/talks/2025/

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
```

Covid Sim (excerpt)

```
/**
 * @brief The basic unit of the simulation and is associated to a geographical location.
 *
 * Interventions (e.g., school closures) are tracked at this level. It contains a list of its
 * members (people), places (schools, universities, workplaces etc.), road networks, links to
 * airports etc.
 */
struct Microcell
{
        /* Note use of short int here limits max run time to USHRT_MAX*ModelTimeStep - e.g. 65536*0.25=16384 days=44 yrs.
           Global search and replace of 'unsigned short int' with 'int' would remove this limit, but use more memory.
        */

        int n; // Number of people in microcell
        int adunit; // admin unit microcell belongs to
        int* members; // array of members/hosts of microcell

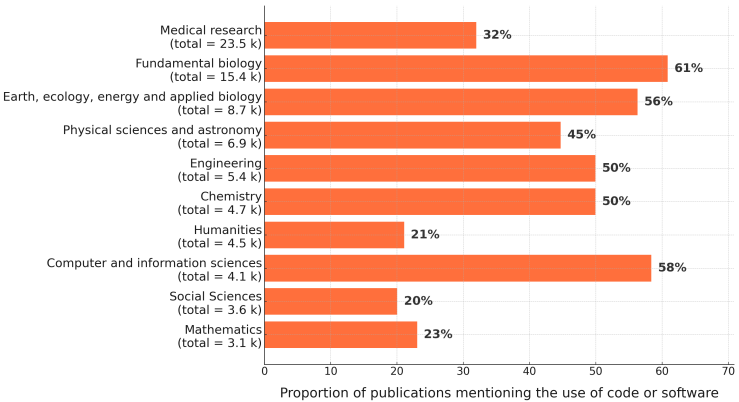
        int* places[MAX_NUM_PLACE_TYPES]; // list of places (of various place types) within microcell
        unsigned short int NumPlacesByType[MAX_NUM_PLACE_TYPES]; // number of places (of various place types) within mircocell
        unsigned short int keyworkerproph, move_trig, place_trig, socdist_trig, keyworkerproph_trig;
        unsigned short int move_start_time, move_end_time;
        unsigned short int place_end_time, socdist_end_time, keyworkerproph_end_time;
        TreatStat moverest, treat, vacc, socdist, placeclose;
        unsigned short int treat_trig, vacc_trig;
        unsigned short int treat_start_time, treat_end_time;
        unsigned short int vacc_start_time;
        IndexList* AirportList;
};
```

Len Shustek, Computer History Museum 2006

"Source code provides a view into the mind of the designer."

(Open) Source Code is a pillar of Open Science

Software powers modern research



French Open Science Monitor 2025

From national to global: Open Science Monitoring Framework Initiative

MINISTÈRE CHARGÉ DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Liberté Égalité Fraternité

unesco



Outline

1. Software (Source Code) as a pillar of (Open) Science
2. Software Heritage at your service
3. European research software through the Software Heritage Very Large Telescope
4. What you can take away this morning

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

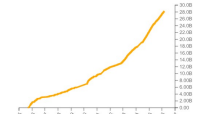
Inria with unesco

The largest open source code archive: one infrastructure, open, shared, non profit

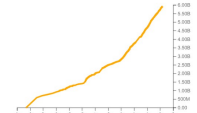
Unique digital common good *built in France since 2015*

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

Source files
28,152,651,759

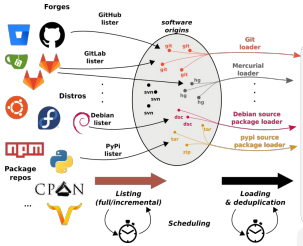


Commits
5,920,787,012



Projects
429,963,770





5000+ platforms

All versions, all history development in a single graph

- **50 × 10⁹** nodes
- **1000 × 10⁹** edges
- ~ **3 PB** of storage

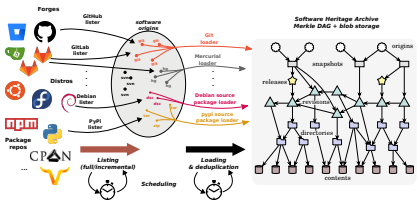
A revolutionary **infrastructure** ensures **availability** guarantees **integrity** enables **traceability**

IBM Microsoft TII Technology Innovation Institute HUAWEI cnrs Infosys cea Ministère de l'Enseignement Supérieur et de la Recherche aws Google Université Paris Cité openinventionnetwork

AdaCore GitHub Université di Pisa Sorbonne Université Université de Lorraine Sant'Anna NWO université PARIS-SACLAY

What we run, for four key needs

Archive (30B+ files, 440M+ projects)



- save now, updateswh, webhooks
- deposit.softwareheritage.org

Describe

- *Intrinsic metadata* from source code
- Contributed the Codemeta generator

Reference (60 billion SWHIDs)

Intrinsic, cryptographically strong IDs



Now in SPDX 2.2, Wikidata

`https://swhid.org` - ISO/IEC 18670

Cite/Credit

- Contributed biblatex-software style
- Software Citation from the archive!

New: research infrastructure, multiannual (French) government support

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE
Liberté Égalité Fraternité

Ouvrir la science !



September 2026
Research Infrastructure with multiannual support
Open Science Strategy 2026–2033, action **SO7.A4**

Nine strong mentions in the Strategy, e.g.:

- record research output with HAL and Recherche Data Gouv *"Software Heritage, a piece of infrastructure with shared governance, plays a major role in the universal preservation of scientific assets."*
- sovereignty: a sovereign forge *"must be complemented by efforts to archive global software assets on Software Heritage"*
- identification measure SO4.A1.M3: *"software (SWHID and DOI), organisations (ROR) and researchers (ORCID)"*

Outline

1. Software (Source Code) as a pillar of (Open) Science
2. Software Heritage at your service
3. European research software through the Software Heritage Very Large Telescope
4. What you can take away this morning

From the **Software Sky** to **Institutional Insight:** Anatomy of an Observational Pipeline

Observation parameters (config file)

Stage 0: The Software Sky

The process begins with a complete map of the software universe, built from the Software Heritage archive, containing all observable software 'stars'. This stage is configured by observation parameters and a spectral filter definition (email domain selection).

Phase 1: Target Selection

Identify the stars associated with a specific institution by selecting contributors using institutional identifiers, which isolates the relevant region of the software sky for focused observation.

Quality Gate: Signal Strength Check

Before proceeding, the pipeline verifies that the selected region contains enough observable signals (software contributions) to produce statistically meaningful results.

Egeait to Nous Kaba

Phase 2: Gravitational Cleaning (Deforking)

The process renvents gravitational artifects such as mitrers and fures from the data to retain only the physically meaningful celestial objects (original software projects).

Phase 4: Spectral Enrichment

Each observed object is euriched with spectral data—such as activits, domains, and licenses—collected from multiple observation instruments to provide a deeper understanding.

Activity

Domains

Licenses

Adeptive Observation Scheduling

This fecture ensures the automatic relation of observation tokens to bypass iassument (DPQ limits, ensuring continuous data collection.

Observation Cache

A oerWixeen observation log is maintained, preventing redandant measurements and speeding up subsequent analyzes.

Phase 5: Astrophysical Analysis

The entiched observations are analysed to compute metrics, classily objects into scientific domains, and generate a synthetic, comprehensive view of the observed system.

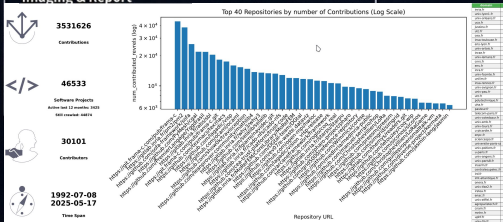
Phase 6: High-Resolution Imaging & Report

3531626
Contributions

46533
Software Projects
Active last 12 months: 3025
Still-created: 16874

30101
Contributors

1992-07-08
2025-05-17
Time Span



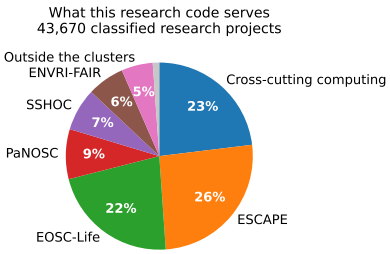
The software of the EU Science Clusters, measured by Software Heritage

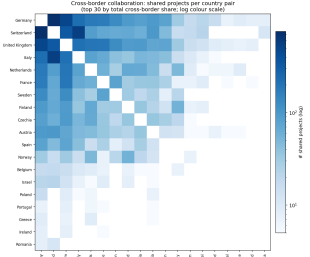
43,670 research code projects, **24,475** contributors, **20** countries.

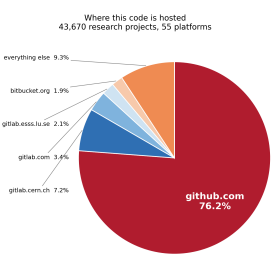
76% of them on **github.com**, 7.2% on CERN's own forge, 2.1% on ESS's.

5.4%, **2,356 projects**, disappeared already, saved in Software Heritage.

the full report







Left: what the code serves; 1.1% carry no cluster label. Middle: 102 country pairs, 6,430 shared projects, over the 43,670 research projects whose contributors resolve to a country. Right: where those projects are hosted; no project on an institutional forge was wrongly admitted.

One archive, all public code: **the very large telescope for source code.**

Strategic autonomy: who can decide that your code stops existing?

The danger: **76%** of research code on **one platform**, one jurisdiction

Forges are not archives

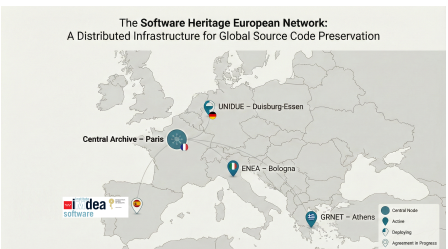
- **platforms close** (Google Code, Gitorious, ... ∼1M projects!)
- **projects change, move, disappear**
- **30 March 2024**: xz-utils accounts suspended, repository unreachable until 10 April (*but available on SWH!*)
- **13 September 2026**: two hours of global GitHub outage

Resilience



Since 2019 **GNU Guix** falls back to the archive **automatically**. **Nix** is next.

Ecumenical digital sovereignty



One non profit open international core infrastructure. Each **mirror** run by an *independent organisation, in its own jurisdiction, on its own stack.*

Outline

1. Software (Source Code) as a pillar of (Open) Science
2. Software Heritage at your service
3. European research software through the Software Heritage Very Large Telescope
4. What you can take away this morning

New: a doctoral course, written to be reused



The Source Code of Science

Archiving, Referencing and Reproducing Research Software. A doctoral course, 89 pages, version 1.6, September 2026.

CC-BY 4.0: yours to reuse, translate and teach from.

Seven chapters and ten appendices, from identifiers to artefact evaluation, with a cookbook, a glossary and a policy matrix.

hal.science/hal-05734828

Selected services we are rolling out

1. Archive your institutional forge(s). We add it to the archive, all of it, permanently. *Nobody to convince: public code needs no permission.*
2. Your own Strategic Software Insights report. *Université Paris Cité, as a Gold member, receives annual editions.*
3. Training and Support. *Morane Gruenpeter oversees this.*
4. Support your OSPO (people that help researches manage their software) *What software is developed in your institution? What licence, discipline, affiliation? The OSPO-Radar product is in full development, you can have your voice heard*

Join the interest groups: it is how institutions get heard

Two Interest Groups

- DIG, the Source Code Deposit Interest Group: for institutions that want a seat where the archive is steered.

  Already in: Université Paris Cité, Università di Pisa, Scuola Normale Superiore, Scuola Superiore Sant'Anna, NWO, …
- ALIG: for archives and research libraries
  Already in: CRKN (seven Canadian universities), the Swiss CSAL, CERN, Couperin, and twenty more (see the members online).

More information: mail to sponsor@softwareheritage.org

(Only/Already) two of the nine Circle U universities are on the list.

Thank you

Roberto Di Cosmo roberto@dicosmo.org

- the archive archive.softwareheritage.org
- archive a repository archive.softwareheritage.org/save
- the course hal.science/hal-05734828
- the annual report
  annex.softwareheritage.org/public/annual-reports/2025/

Want to know more, or join?
Contact Morane Gruenpeter
morane@softwareheritage.org



softwareheritage.org

These slides are CC-BY 4.0.