Software as a Pillar of Open Science

through the Software Heritage global digital infrastructure

Roberto Di Cosmo

Director, Software Heritage
Chair, Software Chapter, CoSO
Inria and Université de Paris Cité

October 1st, 2025

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

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)

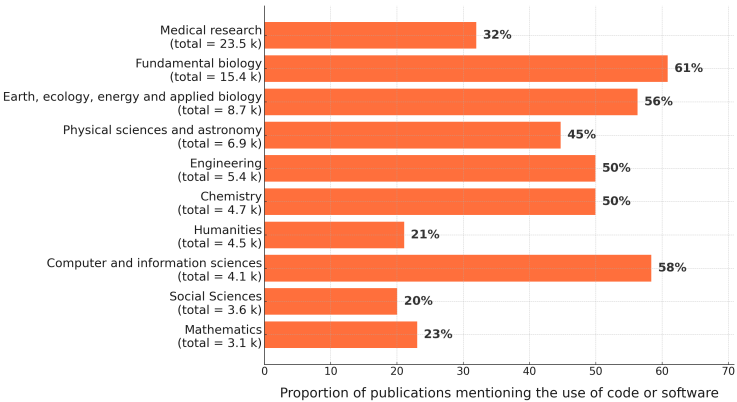
```
/**
 * @brief The basic unit of the simulation and is associated to a geographical location.
 *
 * Interventions (e.g., school closures) are tracked at this level. It contains a list of its
 * members (people), places (schools, universities, workplaces etc.), road networks, links to
 * airports etc.
 */
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

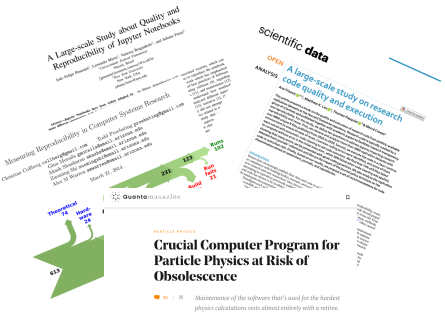
MINISTÈRE CHARGÉ DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Liberté Égalité Fraternité

unesco

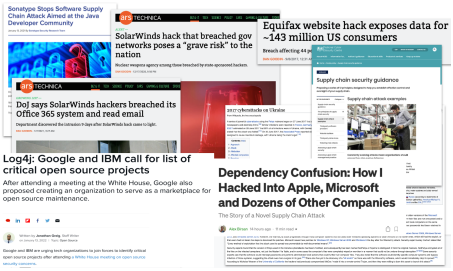
How are we managing our software ?

Reproducibility, maintenance in Academia



(articles: here, here, here and here)

Security, integrity, traceability in Industry



Can they track the software that they

- ship, use, acquire
- has that bug or vulnerability

awareness is raising at the level of public policy

A bird's eye view of arising policies

Publishers

- *"make the software associated to your articles available"* mandates

Artifact Evaluation Committees

- *"review the software artifacts associated to articles"* initiatives (ACM badges, etc.)

Funders

- *"software developed under grant XXXX must be made available under an open license"*

Institutions

- *"add software to the research curriculum"*
- *"software considered a research output (e.g. DFG grant applications)"*
- *"report the software produced from the lab/institute/organization"*

Communities

- *"software must be cited and credited"*

Key needs: Archive, Reference, Describe, Cite and Credit

Archive

Research software artifacts must be properly archived

make sure we can *retrieve* them (*reproducibility*)

Reference

Research software artifacts must be properly referenced

make sure we can *identify* them (*reproducibility*)

Describe

Research software artifacts must be properly described

make it easy to *discover* and *reuse* them (*visibility*)

Cite/Credit

Research software artifacts must be properly cited *(not the same as referenced!)*

to give *credit* to authors (*evaluation*!)

We need a dedicated infrastructure to address these needs: now we have one!

Outline

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

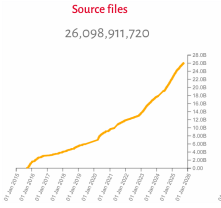
enable analysis of all software source code

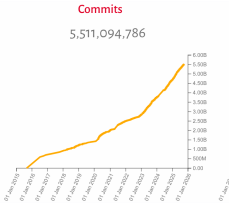
A *universal* software archive, as a shared infrastructure

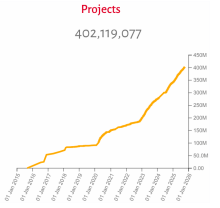
One infrastructure
non profit
open and shared

Unveiled in 2016

The largest archive ever built







Diamond sponsors

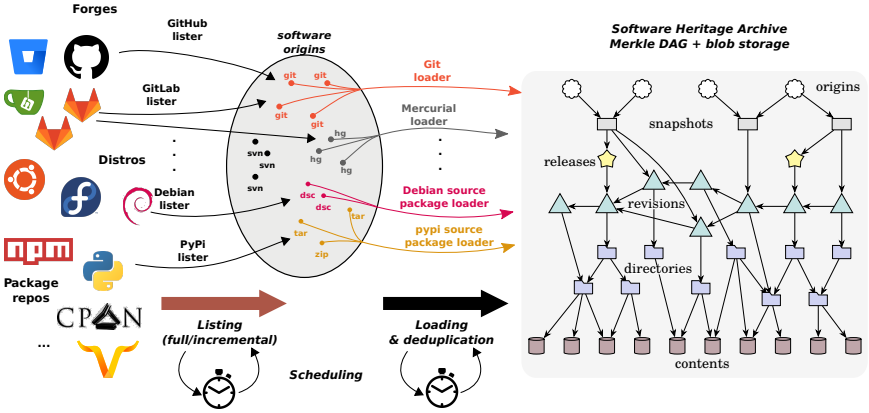
Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors

The archive under the hood

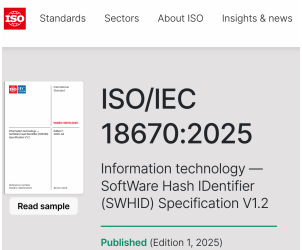


The archive under the hood

Global development history permanently archived in a uniform data model

- over 26 billion unique source files from over 400 million software projects
- ~2PB (compressed) blobs, ~50 B nodes, ~800 B edges

Software Hash Identifiers (ISO 18670)



Criptographically strong identification for *50 billion* software artifacts

Integrity and traceability support reproducibility and transparency

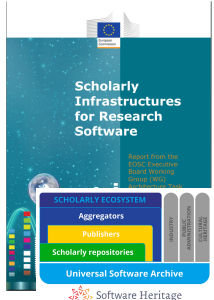
Outline

Demo time!

- Browse + Reference [DIS 18670] (Apollo 11 [excerpt], your work may be already there !)
- Trigger archival, use the updateswh browser extension, configure the webhooks
- Cite from the archive with biblatex-software (CTAN, ACMART)
- 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

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

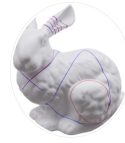
eosc | FAIR-IMPACT
Expanding FAIR solutions across EOSC

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

Outline

Call to action: best practices for ARDC are available… today!

Archive and reference

All source code used in research (*yes, even small scripts!*) *for reproducibility*

- save in Software Heritage
- add SWHID in articles

See detailed HOWTO online

Describe and Cite/Credit

For software you want to put forward (*mention in your CV, reports, etc., get citations and credit for it*), do the following extra steps: video tutorials

- add codemeta.json (see the codemeta generator)
- reference in HAL (*french partners*, see online HAL documentation)
- cite using the biblatex-software package (in CTAN and TeXLive)

- train students and colleagues
- engage journals, conferences, learned societies