Building the software pillar for open science

Roberto Di Cosmo

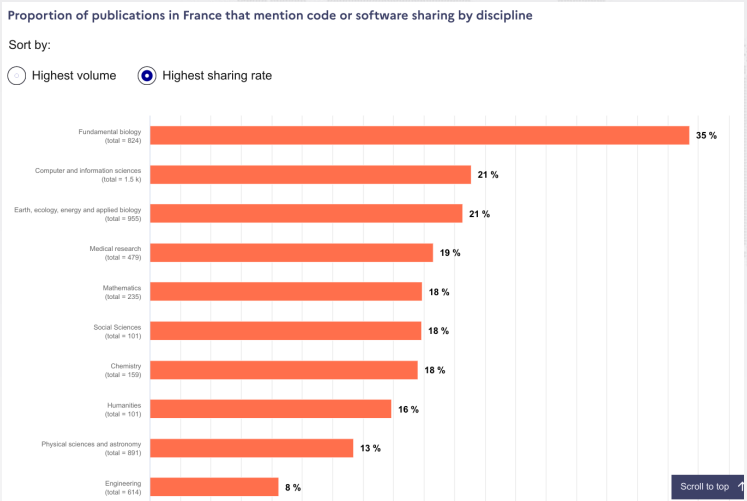
Chair, Software Chapter, National Committee for Open Science
Director, Software Heritage
Inria and Université de Paris Cité

July 16th 2023

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Software in all research areas: the data is now in!



French Ministry, 2023

Open Science Monitor

- 160.000 articles
- *all disciplines*
- 20+% *share* software

Approach:

- Public protocol
- FOSS software
- open data

Knowledge is in the Software *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)

```
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)

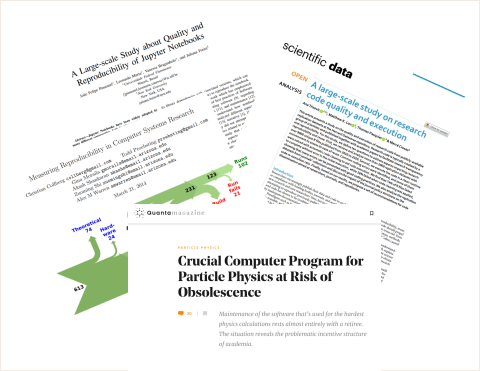
```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    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."

How are we managing our software ?

Availability, Reproducibility, Sustainability, Recognition



(articles: here, here, here and here)

An emerging policy framework

Paris Call Source code 2019, Open Science recommendations 2021, French National Open Science Plan 2021, EOSC SIRS 2020, OSEC 2022, DFG new CV (9/22), NASA Open Science policy (12/22), CERN/NASA summit

What is at stake

ARDC

- Archive for retrieval (*reproducibility*)
- Reference for identification (*reproducibility*)
- Describe for discovery and reuse
- Cite/Credit for credit and evaluation

Before ARDC

- Development practices and tools (VCS, build system, test suites, CI, code quality, …)
- Opening up towards a community (documentation, organization, communication)

Need training, tooling, infrastructures, best practices

Beyond ARDC

- Policies (dissemination, reuse, careers, …)
- Sustainability (legal, financial, etc.)
- Technology transfer
- Advanced technologies and tools (quality, traceability, etc.)

a humbling challenge, a complex one, and *we are all concerned*

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

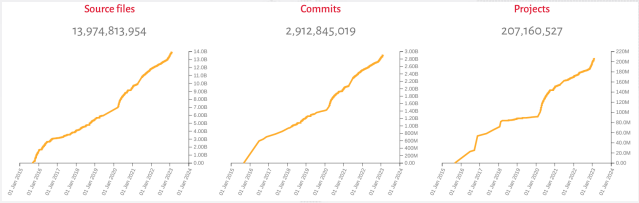
Research infrastructure

enable analysis of all software source code

The largest software archive, a shared infrastructure

One infrastructure
open and shared

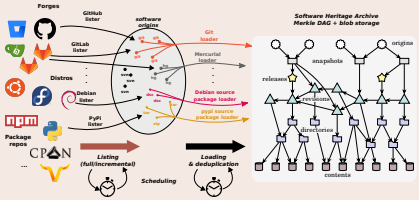
The largest archive ever built



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

Addressing the needs (see ICMS 2020 for details)

Archive (15B+ files, 240M+ projects)



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

Reference (30 billion SWHIDs)

Intrinsic, cryptographically strong IDs

Now in SPDX 2.2, Wikidata

Specification: `https://swhid.org`

Describe

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

Cite/Credit

- Contributed *software citation* style biblatex-software, v 1.2-2 now on CTAN

The floor is yours

We need you: learn, adopt, train, engage, contribute it's a long road, but together we can make it

References


UNESCO, *Draft recommendations on Open Science*
2021, (online)


French Ministry of Research, *Second National Plan for Open Science*
2021, (online)


EOSC SIRS Task Force, *Scholarly Infrastructures for Research Software*
2020, Publications office of the European Commission, (10.2777/28598)


R. Di Cosmo, *Archiving and Referencing Source Code with Software Heritage*
International Conference on Mathematical Software 2020 (10.1007/978-3-030-52200-1_36)


J.F. Abramatic, R. Di Cosmo, S. Zacchiroli, *Building the Universal Archive of Source Code*
CACM, October 2018 (10.1145/3183558)