Building Infrastructures for Open Source Software

a short introduction

Roberto Di Cosmo
UN Open Source Retreat 2024

Director, Software Heritage
Inria and Université de Paris Cité

May 14th 2024

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- *35+ years* of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- *25+ years* of Free and Open Source Software
- *15+ years* building and directing structures for the common good

1999 *DemoLinux* – first live GNU/Linux distro

2007 *Free Software Thematic Group*
150 members 40 projects 200Me

2008 *Mancoosi project* www.mancoosi.org

2010 *IRILL* www.irill.org

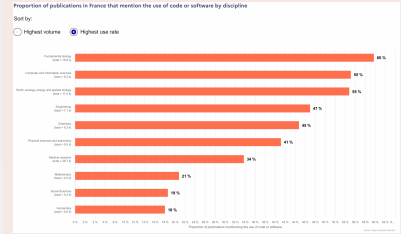
2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

2021 *EOSC Task Force on Infrastructures for Software*, European Union

Software is a pillar of Open Science

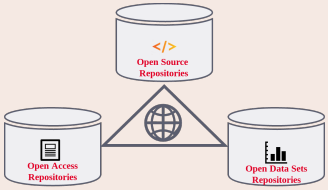
Software powers modern research



Over 20% of articles using software across all disciplines share it

2024 French Open Science Monitor

Key pillar: software



Links are important

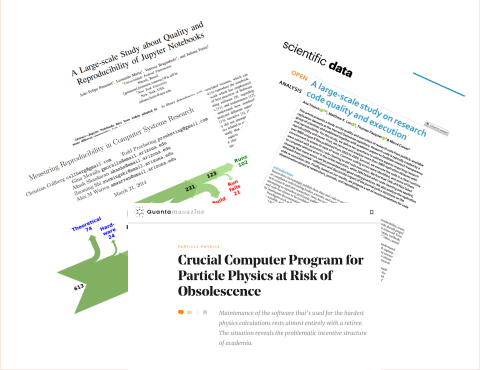
Nota Bene

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

access to the *source code* is essential!

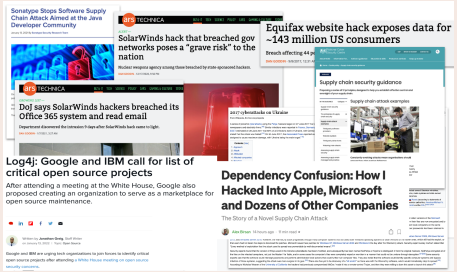
How are we managing our (open source) software ?

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

Outline

French National plan for Open Science, 2021-2024

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

Liberté
Égalité
Fraternité

SECOND FRENCH PLAN FOR OPEN SCIENCE

Generalising open science in France 2021-2024

1

Launch on 6 July 2021 by Frédérique Vidal, Minister for Higher Education, Research and Innovation

- Multiplying the **levers for change** in order to **generalise open science practices**
- Structuring the **policy for opening up or sharing research data**
- New commitments to the **opening of source code** produced by research
- **European and international inclusion** in the context of the French Presidency of the European Union
- **Disciplinary and thematic variations**: open science policies must be adapted to disciplinary specificities

2

Path Three :

Opening up and promoting source code produced by research

7 **Recognize and support the dissemination under an open source license of software** produced by publicly funded research programmes

8 **Highlight the production of source code** from higher education, research and innovation

9 Define and promote an **open source software policy**

« The opening of software source code is a major challenge for the **reproducibility** of scientific results. »

« Distribution of software products under **open source licence** will be preferred. »

3

Define and promote an open source software policy

- Produce a **National Charter for Open Source Software** coming from higher education, research and innovation
- Develop the **link between data and software** through a network of **Chief Data Officers** in the various universities and research performing organisations.
- Develop the **economic models of open source software** and make them known within commercialization services
- **Support Software Heritage** and recommend it for the archiving and referencing of source code

Recognise source code as a contribution to research

- Create an **open source research software prize**
- **Provide greater recognition** for software production in the career of researchers, research support staff

Build an ecosystem that connects code, data and publications

- Develop **proper coordination** between software forges, open publication archives, data repositories and the scientific publishing sector.

4

Software on par with research articles and data

Software College in the National Committee for Open Science

- Five action lines (licence is just the tip of the iceberg, see details online)
- Co-Chairs: Roberto Di Cosmo, François Pellegrini
- Members: 20+ from all research fields
- Key productions:
  - Source Code primer 
  - Report on software forges (needs, limitations, options): 
    - Address GitLab instance limitations for global contribution: https://code.gouv.fr/fr/bluehats/outils-de-forge-2024/
  - *National research software award.* Open Research Europe 2023 

French National plan for Open Science awards 2022 and 2023

First edition, 2022 prize

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Accueil > Recherche > Science ouverte

Publié le 05.02.2022

Sommaire

- The Coq proof assistant : lauréat de la catégorie Scientifique et technique
- Scikit-learn : lauréat de la catégorie Communauté
- Faust : lauréat de la catégorie Documentation
- Gammapy : prix du jury
- Jury

Remise des prix science ouverte du logiciel libre de la recherche

Le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation remet pour la première année les Prix science ouverte du logiciel libre de la recherche. Dix logiciels mis au point par des équipes françaises sont récompensés pour leur contribution à l'avancée de la connaissance scientifique.

- 129 projects
- 4 awards
- 6 accessit
- first edition

- Coq proof assistant
- Scikit-Learn ML/AI
- Faust music
- Gammapy astronomy

Second edition, 2023 prize

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Accueil > Recherche > Science ouverte

Publié le 29.11.2023

Sommaire

- PlanQOOLN : espoir de la catégorie « Scientifique et technique »
- Smilei : lauréat de la catégorie « Scientifique et technique »
- NoiseCapture : espoir de la catégorie « Communauté »
- OCaml : lauréat de la catégorie « Communauté »
- ReOps : espoir de la catégorie « Documentation »
- Brian : lauréat de la catégorie « Documentation »
- Fink : espoir de la catégorie « Coup de cœur » du jury
- HyphE : lauréat de la catégorie « Coup de cœur » du jury
- Jury

Remise des prix science ouverte du logiciel libre de la recherche 2023

Le ministère de l'Enseignement supérieur et de la Recherche remet pour la deuxième édition les Prix science ouverte du logiciel libre de la recherche. Huit logiciels développés par des équipes françaises sont récompensés pour leur contribution à l'avancée de la connaissance scientifique ou pour le caractère prometteur de leurs travaux.

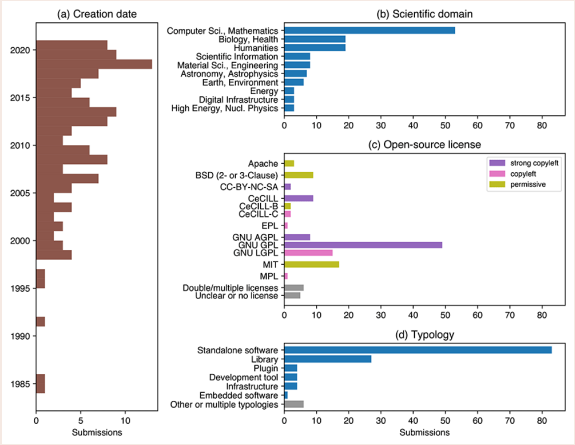
- 66 projects
- 4 awards
- 4 "espoirs"
- will be run annually

Blueprint and data from the first edition (2021-2022)

Blueprints and analysis available

- goals, design decisions
- challenges and solutions
- lessons learned
- detailed data

A look at the data



Emulation is working

Australia: here and here; Germany: Helmholtz; European Commission: a CSA; …

A few key shared needs across all areas

We need a universal shared infrastructure to address them!

Outline

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

An international, non profit initiative built for the long term

Sharing the vision

United Nations Educational, Scientific and Cultural Organization

And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Diamond sponsor

Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors

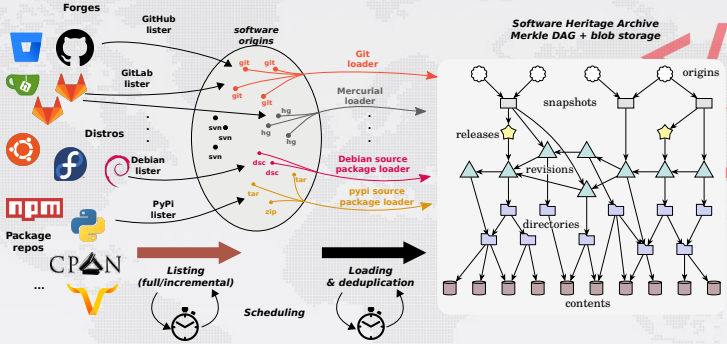
The largest software archive, a shared infrastructure

One infrastructure
open and shared

The largest archive ever built

Bitbucket		git
2,509,402 origins	56,983 origins	24,600 origins
26,599 origins	debian 136,338 origins	53,297 origins
GitHub	gitiles	GitLab
197,883,004 origins	10,171 origins	4,216,298 origins
git	Gogs	GO
2,926 origins	172 origins	971,549 origins
Guix	GNU	heptapod
14,482 origins	354 origins	1,207 origins
launchpad	Maven	NixOS
503,631 origins	312,461 origins	14,482 origins

Address common Open Science and Open Source needs: archival



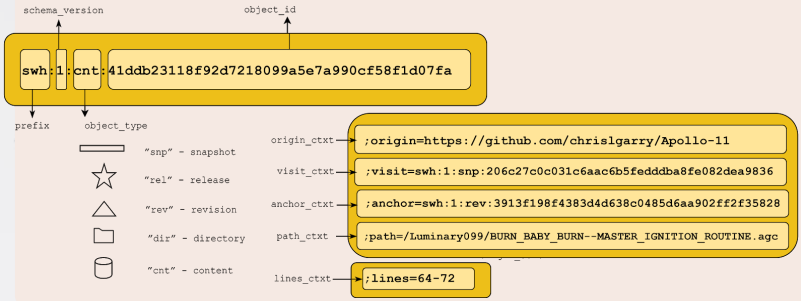
Global development history permanently archived in a uniform data model

- over 18 billion unique source files from over 290 million software projects
- ~1.5PB (compressed) blobs, ~35 B nodes, ~500 B edges

Address common Open Science and Open Source needs: reference

Software Heritage Identifiers (SWHID)

link to full docs



40+B
intrinsic,
decentralised,
cryptographic

Full fledged *source code references* for traceability, integrity and reproducibility

- Linux Foundation SPDX 2.2
- IANA-registered "swh:"
- WikiData property P6138

Examples: Apollo 11 AGC excerpt, Quake III rsqrt
Guidelines available, see the HOWTO

Breaking news: standardisation, see swhid.org

Outline

A walkthrough

- Browse and Reference (e.g. Apollo 11 [excerpt], your work may be already there !)
- Trigger archival, use the updateswh browser extension, configure the webhooks
  - example: French public open source at: code.gouv.fr
- 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

Outline

Institutional Software Portal

DALL-E's view of an instutional portal

Problem: institutions need to identify projects of interest…
… projects spread across 100s of platforms

Solution: personalised institution portals on SWH

- deposit and curation of (meta)data
- presentation
- extraction

Why Software Heritage:

- platform agnostic, metadata from multiple institutions
- shared knowledge base

Certified deposit

DALL-E's view of SWH as CD

Problem: organizations need trusted timestamps on software deposits…

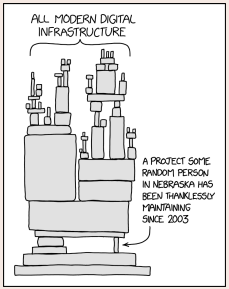
Solution: personalised deposit portals with a secure timestamp conforming to ISO/IEC 18014

Why Software Heritage:

- SWH's Merkle tree is *half* a blockchain
- SWHID guarantees integrity
- universal archive, standardised
- minimal disruption to current deposit workflow

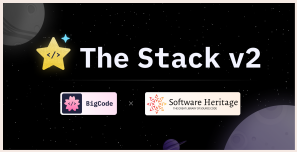
Cybersecurity, AI, Tech, Science …

Cybersecurity: tracing vulnerabilities



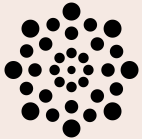
See the SWHSec project

(Generative) AI



The Stack V2 (subset of Software Heritage) used to build StarCoder2, *best open AI model for coding today* (3, 7 and 15B parameters)

Big Data challenges



- big data infrastructure
- efficient queries
- integration with other knowledge graphs

Outline

A rally flag for a grand vision

Bring together academia, industry, governments, communities

"to build a reference, global infrastructure for open and better software"

Software Heritage is the first brick ...

- vendor neutral
- open source
- a worldwide initiative
- a long term initiative

... that will enable

- archival, reference, integrity
- qualification, sharing and reuse
- a global software knowledge base
- test and deploy world class tooling

A lot more is needed

Software Heritage can be the *catalyser* of a way bigger undertaking

The floor is yours

Let's work together!

Annual report

5 years in 5 minutes

Link

Evolution of our codebase

Link