Open Source challenges and opportunities

from Open Science to the Supply Chain

Roberto Di Cosmo
Congrès de la SIF

Director, Software Heritage
Inria and Université de Paris Cité

April 6th 2023

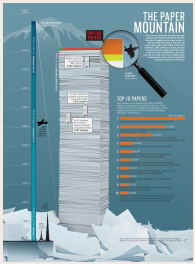
Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

Software is a pillar of Open Science

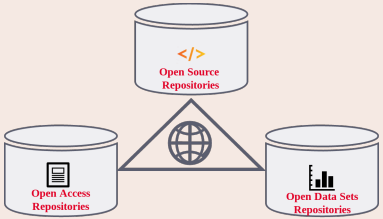
Software powers modern research



[…] software […] essential in their fields.
Top 100 papers (Nature, 2014)

Sometimes, if you dont have the software, you dont have the data
Christine Borgman, Paris, 2018

A key pillar: software (source code)



The links in the picture are important

Nota Bene

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*

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)

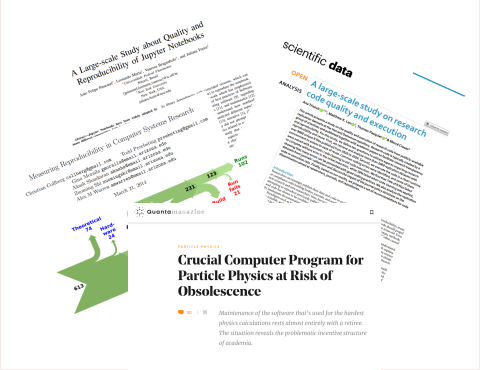
```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    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."

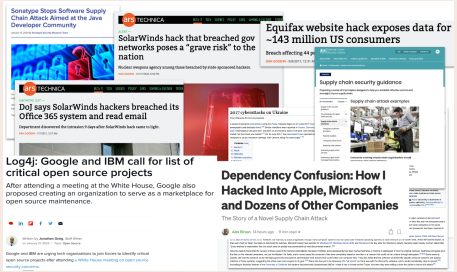
How are we managing our software ?

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

Outline

French National plan for Open Science, 2021-2024

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

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 **reproductibility** of scientific results. »

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

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

A few basic needs for software in Open Science

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*!)

These are also industry needs!

Open Source is growing…

Software is eating the world

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts

ESSAY

Why Software Is Eating The World

By Marc Andreessen

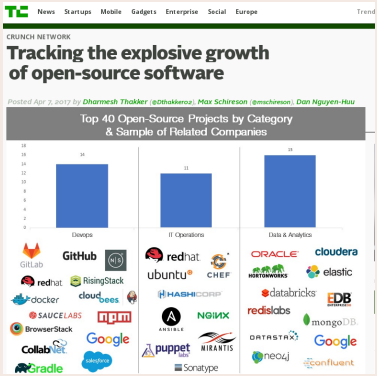
August 20, 2011

This week, Hewlett-Packard (where I am on the board) announced that it is exploring jettisoning its struggling PC business in favor of investing more heavily in software, where it sees better potential for growth. Meanwhile, Google plans to buy up the cellphone handset maker Motorola Mobility. Both moves surprised the tech world. But both moves are also in line with a trend I've observed, one that makes me optimistic about the future

Software companies outperform
or buy out traditional companies

Marc Andreesen, 2011

Open Source is eating the Software World



TC News Startups Mobile Gadgets Enterprise Social Europe Trend

CRUNCH NETWORK

Tracking the explosive growth of open-source software

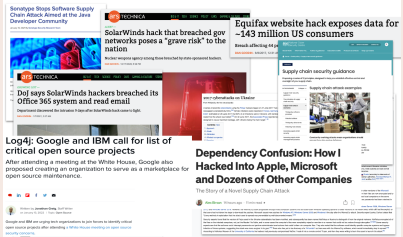
Posted Apr 7, 2017 by Dharmesh Thakker (@dthakkerca), Max Schireson (@mschireson), Dan Nguyen-Huu

Top 40 Open-Source Projects by Category & Sample of Related Companies

Reuse is the new rule

80% to 90% of a new application is … just reuse! (Sonatype survey, 2017)

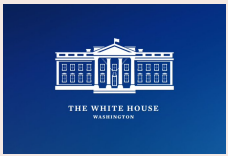
Software supply chain attacks abound



Can you track the software that…

- you ship
- you use
- you acquire
- has that bug
- has that vulnerability

KYSW: Know Your SoftWare - like KYC in banking



Sec. 4. Enhancing Software Supply Chain Security

ensuring and attesting, to the extent practicable, to the integrity and provenance of open source software

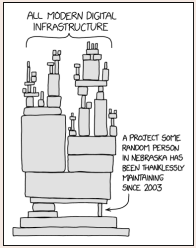
May 2021 POTUS Executive Order

Can we fulfil together these shared needs?

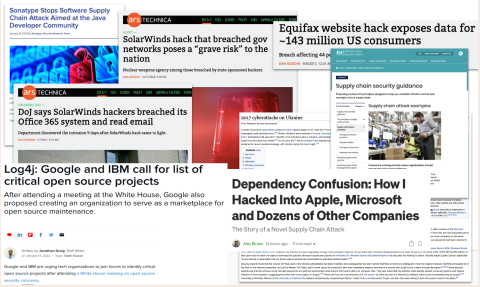
Outline

Software supply chain and its issues

Complex digital infrastructure



Software supply chain in the news

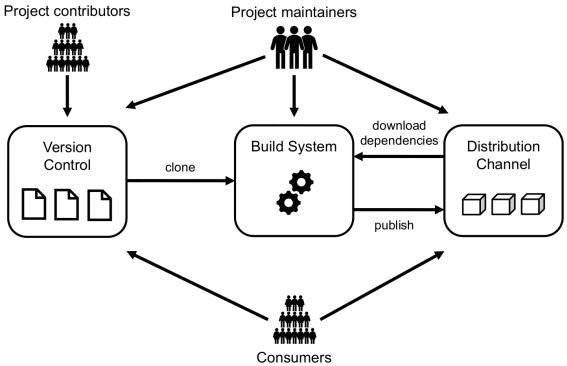


Software Supply Chain attacks

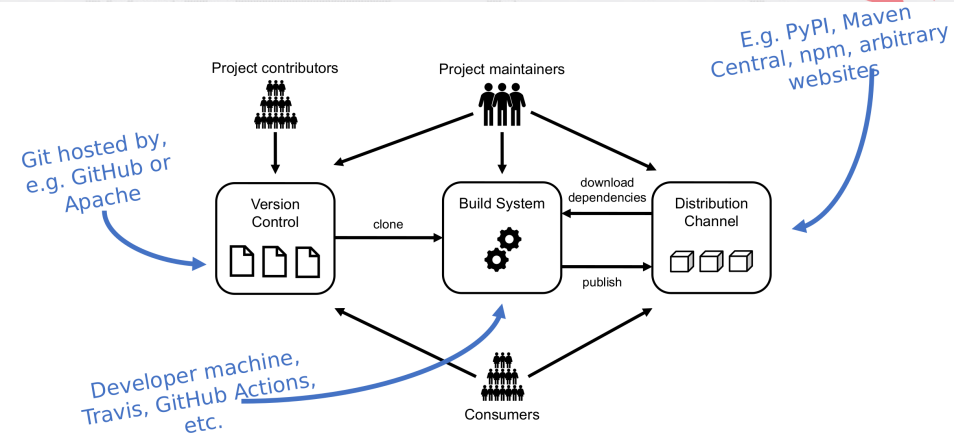
Malicious code injection into software components to compromise downstream users

- March 2022 node-ipc and peacenotwar (CVE-2022-23812)
- Dec 2021 Apache Log4j Remote Code Execution (Log4Shell, CVE-2021-44228)
- Nov 2018 Attack on NPM package event-stream

Software supply chain in a picture



Software supply chain in a picture



A long road ahead

Vertical approach

improve security of *each component* separately

Horizontal approach

explore *the whole supply chain*

A few key challenging properties

findability	needs qualified metadata
availability	needs an archive and a system of identifiers
integrity	needs crypto
traceability	needs a global provenance database
reproducibility	needs groundbreaking tools

We need a *global coordinated effort…*
and a *common, open, shared* infrastructure to track *all (Open Source) software*!

Forges are key platforms, but they are *not* enough!

2015: the first big bad news

Google Code and Gitorious.org shutdown: ~1M endangered repositories

- broken links in the web of knowledge (my papers too)

Big bad news keep coming in

- summer 2019: BitBucket announces Mercurial VCS sunset
- july 2020: BitBucket erases *250.000+* repositories (including research software)
- summer 2022: GitLab.com considers erasing all projects that are inactive for a year

In Academia too!

- 2021: Inria's old gforge is unplugged… breaks the Opam build chain for OCaml

source code is spread across hundreds of them…

lack of uniformity, no persistence guarantee

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

One infrastructure
open and shared

Largest archive

Technology

- transparency and FOSS
- replicas all the way down

Content (billions!)

- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- multi-stakeholder

Sharing the vision

United Nations Educational, Scientific and Cultural Organization

And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Inria

Diamond sponsor

cea

Platinum sponsors

cnrs, SOCIETE GENERALE, intel, Microsoft, MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE, HUAWEI

Gold sponsors

openinventionnetwork, SORBONNE UNIVERSITÉ, Université de Paris

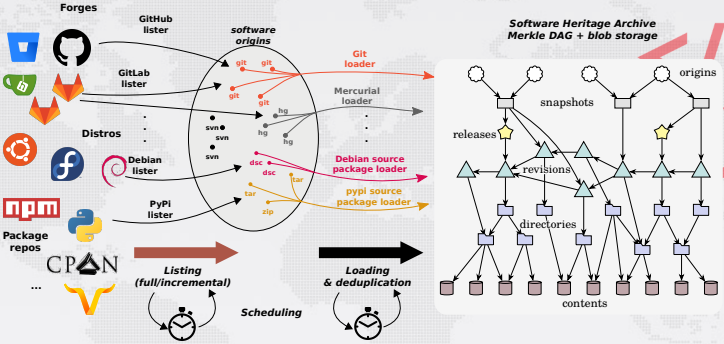
Silver sponsors

AdaCore, MINISTÈRE DES ARMÉES, DGA, Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE, GitHub, Google, UNIVERSITÀ DI PISA

Bronze sponsors

Red Hat, SCANOSS, SCUOLA NORMALE SUPERIORE

A peek under the hood: a universal archive



Global development history permanently archived in a uniform data model

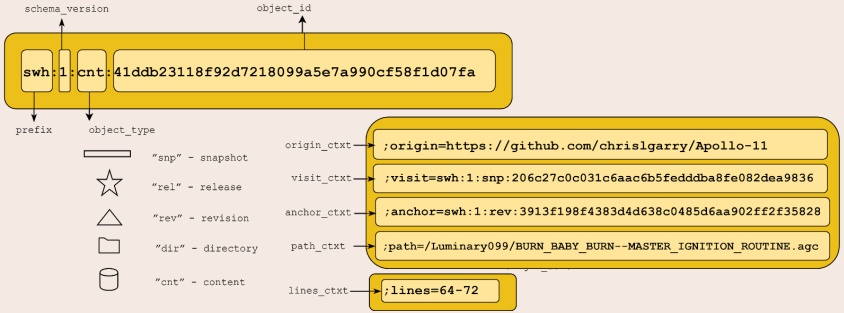
- over 14 billion unique source files from over 210 million software projects
- ~1PB (compressed) blobs, ~30 B nodes, ~400 B edges

Intrinsic Identifiers for software artefacts

Software Heritage Identifiers (SWHID)

link to full docs

25+B intrinsic, decentralised, cryptographically strong identifiers, SWHIDs



Emerging standard : Linux Foundation SPDX 2.2; IANA registered; WikiData P6138

Full fledged *source code references* for reproducibility

Examples: Apollo 11 AGC excerpt, Quake III rsqrt; Guidelines available, see ICMS 2020

A quick tour

- Browse (e.g. Apollo 11, and your work may be already there !)
- Trigger archival, use the updateswh browser extension (GitHub action available too)
- Get and use SWHIDs (full specification available online)
- Cite software with biblatex-software package from CTAN
  - Overleaf ACMART template available
- Example in journals: article from IPOL
- Example with Parmap: devel on Github, archive in SWH, curated deposit in HAL
- Extracting all the software products for Inria, for CNRS, for CNES, for LIRMM or for Rémi Gribonval using HalTools
- Curated deposit in SWH via HAL, see for example: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, ...
- Example use in research articles:
  - compare Fig. 1 and conclusions in the 2012 version and the updated version
  - SWHID in a replication experiment

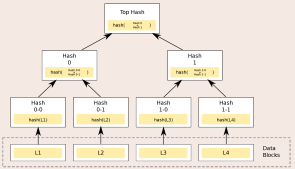
A revolutionary infrastructure

The *graph* of Software Development

All software development in a single graph …

- enable traceability

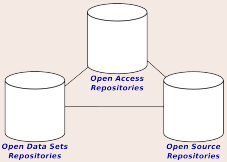
The *blockchain* of Software Development



… a Merkle graph

- ensure integrity

A *pillar* of Open Science



Reference archive of Research Software

- reproducibility
- reference

Reference platform for *Big Code*

uniform data structure

- large scale studies
- machine learning, AI, …

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

You can help!

fund and/or *develop* SWH, *use* SWH research, *build* tools, contribute to swhid.org

The floor is yours

Let's all work together!

Questions?

References


R. Di Cosmo, *A revolutionary infrastructure for Open Source*, 2021, EU Software Forum (slides) (video)


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


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)