Software Heritage

A revolutionary infrastructure for Open Source

Roberto Di Cosmo
Director

November 24th, 2021
TechWeek Société Générale

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Introduction
2. Knowing Open Source Software
3. The Software Heritage initiative
4. Demo time!
5. The road ahead

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- *30 years* of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- *20 years* of Free and Open Source Software
- *10 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

Outline

1. Introduction
2. Knowing Open Source Software
3. The Software Heritage initiative
4. Demo time!
5. The road ahead

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

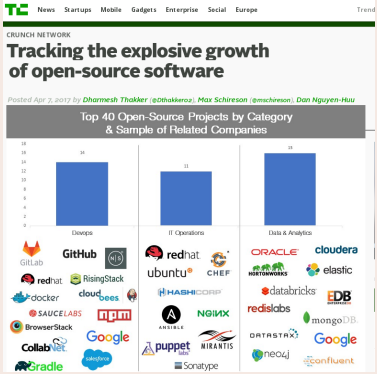
August 20, 2011

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 (@Dthakkeroz), Max Schireson (@mschireson), Dan Nguyen-Hiu

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)

Where does reused software come from?

Do *you* know where it comes from?

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

KYSW: Know Your SoftWare

Like KYC in banking, KYSW is now essential all over IT…

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

… concerns are growing too, cont'd

Several key issues

- *findability* : what code is relevant for that application? needs qualified metadata
- *availability* : can we get this code now? tomorrow? in 20 years? from that country? need an archive and a system of identifiers
- *integrity* : is this really the *right* code? needs crypto
- *traceability* : where are the parts of this system coming from? needs a global provenance database
- *reproducibility* : can we reliably rebuild everything? needs groundbreaking tools
- *traceability*, *safety*, *security*, etc.

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

Outline

1. Introduction
2. Knowing Open Source Software
3. The Software Heritage initiative
4. Demo time!
5. The road ahead

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 all software source code

Research infrastructure

enable analysis of all software source code

Cultural Heritage | Industry | Research | Public Administration

Software Heritage

Technology

- transparency and FOSS
- replicas all the way down

Content (billions!)

- intrinsic identifiers
- facts and provenance

Organization

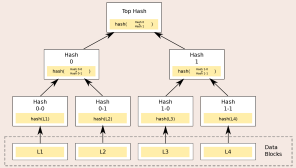
- non-profit
- multi-stakeholder

A revolutionaly infrastructure for open source

The *graph* of Software Development

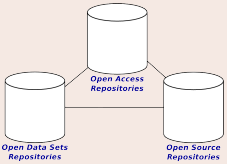
All software development in a single graph …

The *blockchain* of Software Development



… a single Merkle graph!

A *pillar* of Open Science



Reference archive of Research Software

Reference platform for *Big Code*

A single, uniform data structure

Software Heritage Identifiers (SWHID)

link to full docs

An emerging standard

- in Linux Foundation's SPDX 2.2
- IANA registered, WikiData property P6138

Examples:

- Apollo 11 AGC excerpt,
- Quake III rsqrt

Contributing to cybersecurity

Key features

- neutral, common, shared, open, non-profit, *reference knowledge base*
- stores all the source code and its development history, providing a *transparent source of trust*
- cryptographically strong SWHIDs to *track source code artifacts* (file, directory, revision, release, snapshot)

Software Heritage

- ensures availability
- guarantees integrity
- enable traceability

Outline

1. Introduction
2. Knowing Open Source Software
3. The Software Heritage initiative
4. Demo time!
5. The road ahead

A walkthrough

- Browse the archive
- Trigger archival of your preferred software in a breeze
- Get and use SWHIDs (full specification available online)
- Cite software with the biblatex-software style from CTAN
- Example use in a research article: compare Fig. 1 and conclusions
  - in the 2012 version
  - in the updated version using SWHIDs and Software Heritage
- Example use in a research article: extensive use of SWHIDs in a replication experiment
- Curated deposit in SWH via HAL, see for example: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, ...
- Rescue landmark legacy software, see the SWHAP process with UNESCO

Industry use cases (selection)

Open Source complete and corresponding source code distribution (Intel)

Software Heritage members can:

- archive source code in Software Heritage, distribute only the SWHID

Traceability and integrity (OIN for the *Linux System Definition*)

Software Heritage members can:

- archive source code in Software Heritage
- track it and verify its integrity using its SWHID

And much more!

- compliance (collaborations with Intel, FossId, CAST, …)
- security (ongoing collaboration, US Department of Commerce)
- supply chain management, long term archive *add your use case here*

Outline

1. Introduction
2. Knowing Open Source Software
3. The Software Heritage initiative
4. Demo time!
5. The road ahead

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

Platinum sponsors

Gold sponsors

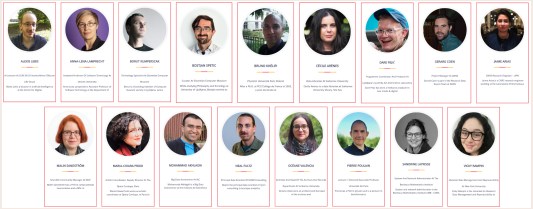
Silver sponsors

Bronze sponsors

A growing community

Team

Ambassadors



Foundations and grantees

- Castalia, CottageLabs
- EasterEggs, OcamlPro
- Octobus, Sperling, Tweag.io

Mirrors and storage partners

"Let us save what remains: … by such a multiplication of copies, as shall place them beyond the reach of accident."

— Thomas Jefferson

Enea, FossId

CEA, RedHat

You can help

- use Software Heritage and SWHIDs in your KYSW process
- become a Software Heritage Ambassador
- spread the word, help adoption, bring new partners

Breaking news: *tuesday 30/11/2021*

- follow the UNESCO public event for 5 years of Software Heritage
- contribute to the #givingtuesday crowdfunding campaign on @swheritage

Questions?

Library of Alexandria of code

- recover the past
- structure the future

A CERN for Software

- build better software
  - for industry
  - for society as a whole