The Software Commons — Production, risks, and opportunities

Stefano Zacchiroli

Télécom Paris, Polytechnic Institute of Paris
stefano.zacchiroli@telecom-paris.fr

23 June 2022
The Materiality of Intangible Goods
Jena, Germany

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

About the speaker

- Professor of Computer Science, Télécom Paris, Institut Polytechnique de Paris
- Free/Open Source Software activist (20+ years)
- Debian Developer & Former 3x Debian Project Leader
- Former Open Source Initiative (OSI) director
- Software Heritage co-founder & CTO

Software is everywhere (and a key mediator) in society

Outline

The Commons and FOSS

Definition (Commons)

The commons is the cultural and natural resources accessible to all members of a society, including natural materials such as air, water, and a habitable earth. These resources are held in common, not owned privately.

Definition (Software Commons)

The software commons consists of all computer software which is available at little or no cost and which can be altered and reused with few restrictions. Thus *all open source software and all free software are part of the [software] commons.* [...]

Software *source code* is precious human 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."

But *where* is this commons?

- many disparate development platforms, with a few dominant players (e.g., GitHub)
- a myriad places where distribution may happen
- most of them operated by for-profit companies

Software source code is fragile

Like all digital information, FOSS is fragile

- link rot: projects are created, moved around, removed
- business-driven code loss (e.g., Gitorious, Google Code, Bitbucket)
- data rot: physical media with legacy software decay

If a website disappears you go to the Internet Archive…

where do you go if (a repository on) GitHub or GitLab goes away?

Takeaways

Software Commons

- Good news: thanks to FOSS developers we are growing by the day a novel kind of digital commons rich in valuable human knowledge: the software commons.

Risk of losing it

- However: there is a risk of "ruining" it (specifically: losing access to it forever).
- As per other commons (including traditional material commons, like natural resources), we need to apply individual care, public policies, and proactive initiatives to mitigate this risk.

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 all software source code

Research infrastructure

enable analysis of all software source code

The largest public source code archive, principled

bit.ly/swhpaper

archive.softwareheritage.org

Technology

- transparency and FOSS
- replicas all the way down

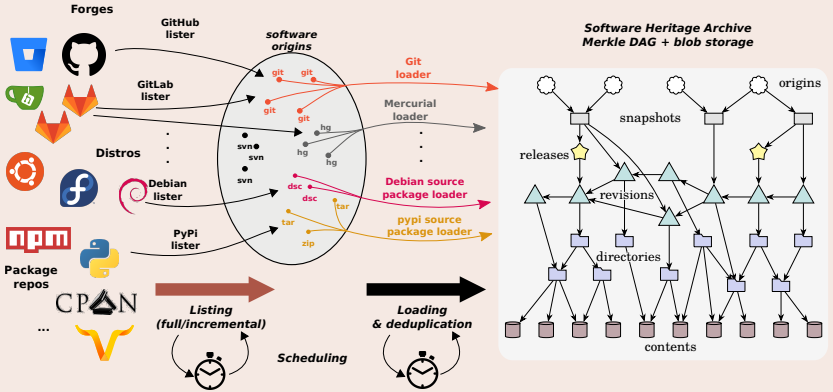
Content (billions!)

- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- multi-stakeholder

A peek under the hood: a global view on the software commons

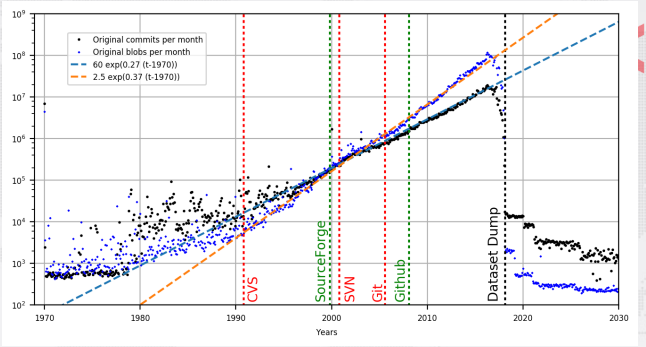


A global graph linking together fully deduplicated source code artifact (files, commits, directories, releases, etc.) to the places that distribute them (e.g., Git repositories), providing a unified view on the entire *Software Commons*.
(Size: ~30 B nodes, ~300 B edges, ~1 PiB blobs)

Outline

The Software Commons is growing fast and becoming more diverse

An observatory on the Software Commons enables studying broad phenomena and trends.



- Original content released as public code doubles every 22-30 months (Di Cosmo, Rousseau, Zacchiroli. Emp.Sw.Eng. 2020).
- Other examples: long-term population trends such as gender diversity (Zacchiroli, IEEE Software 2021) and geography diversity (Rossi and Zacchiroli. ICSE & MSR 2022) in public code.

Who produces Free/Open Source Software?

- Historically, a community of hobbyists motivated by liberating computer users
- Today: the above + independent professionals + a lot of for-profit companies, producing FOSS they control in various ways (licensing, governance, etc.)
- But what is the ratio? We lack *comprehensive* overviews of the landscape. Challenges:
  - Size
  - What do you measure? (projects? SLOCs? popularity?)
  - Free riding is not always visible
- How much exploitation (of labor) and appropriation (of value) is happening?

Mathieu O'Neil, Xiaolan Cai, Laure Muselli, Fred Pailler, Stefano Zacchiroli
The Coproduction of Open Source Software by Volunteers and Big Tech Firms
Digital Commons Policy Council, 2021.

Mathieu O'Neil, Laure Muselli, Xiaolan Cai, Stefano Zacchiroli
Co-producing industrial public goods on GitHub: Selective firm cooperation, volunteer-employee labour and participation inequality
New Media & Society, April 2022.

Outline

Outlook

Takeaways

- Collectively, free/open source software developers are building a novel form of digital commons: the software commons.
- It benefits society as a library of technical knowledge and is exploited by for-profit companies as an industrial public good.
- The risk of losing access to it should be countered with *individual care, public policies, and proactive initiatives*.

Open questions / discussion leads

- Who is producing the software commons? (volunteers/employees/civil servants/…)
- Who is benefiting the most from the software commons?
- Do we need different public policies and/or licensing strategies to fix any of this?
- What other risks is the software commons facing?