Preserving and Enriching the Fabric of Knowledge

Software Heritage as a Global Digital Infrastructure for Open Science and Transparency in AI

Roberto Di Cosmo

Director, Software Heritage
Inria and Université de Paris Cité

May 15th, 2025

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 *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
```

Covid Sim (excerpt)

```
/**
 * @brief The basic unit of the simulation and is associated to a geographical location.
 *
 * Interventions (e.g., school closures) are tracked at this level. It contains a list of its
 * members (people), places (schools, universities, workplaces etc.), road networks, links to
 * airports etc.
 */

struct Microcell
{
        /* Note use of short int here limits max run time to USHRT_MAX*ModelTimeStep - e.g. 65536*0.25=16384 days=44 yrs.
           Global search and replace of 'unsigned short int' with 'int' would remove this limit, but use more memory.
        */

        int n; // Number of people in microcell
        int adunit; // admin unit microcell belongs to
        int* members; // array of members/hosts of microcell

        int* places[MAX_NUM_PLACE_TYPES]; // list of places (of various place types) within microcell
        unsigned short int NumPlacesByType[MAX_NUM_PLACE_TYPES]; // number of places (of various place types) within mircocell
        unsigned short int keyworkerproph, move_trig, place_trig, socdist_trig, keyworkerproph_trig;
        unsigned short int move_start_time, move_end_time;
        unsigned short int place_end_time, socdist_end_time, keyworkerproph_end_time;
        TreatStat moverest, treat, vacc, socdist, placeclose;
        unsigned short int treat_trig, vacc_trig;
        unsigned short int treat_start_time, treat_end_time;
        unsigned short int vacc_start_time;
        IndexList* AirportList;
};
```

Len Shustek, Computer History Museum 2006

"Source code provides a view into the mind of the designer."

The Paris Call on Software Source code (2019, UNESCO)

Experts call for greater recognition of software source code as heritage for sustainable development

.6 November 2018

UNESCO, Inria, Software Heritage invite

40 international experts meet in Paris …

The call is published on February 2019

"Recognise software source code as a fundamental enabler in all aspects of human endeavour"

Unveiled in 2016

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

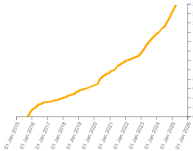
A *universal* software archive, as a shared infrastructure

One infrastructure
open and shared

The largest archive ever built

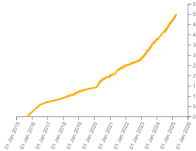
Source files

23,828,196,855



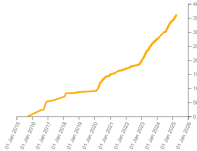
Commits

5,006,891,310



Projects

362,987,832



Directories

18,809,563,175

Authors

90,221,599

Releases

106,890,095

Diamond sponsors

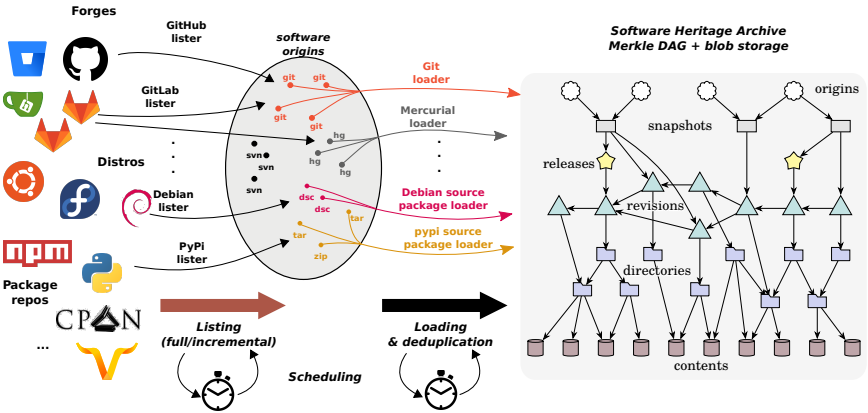
Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors

The archive under the hood



The archive under the hood

Global development history permanently archived in a uniform data model

- over 23 billion unique source files from over 360 million software projects
- ~2PB (compressed) blobs, ~50 B nodes, ~800 B edges

Software Hash Identifiers (ISO 18670)

Standards Sectors About ISO Insights & news

Read sample

ISO/IEC 18670:2025

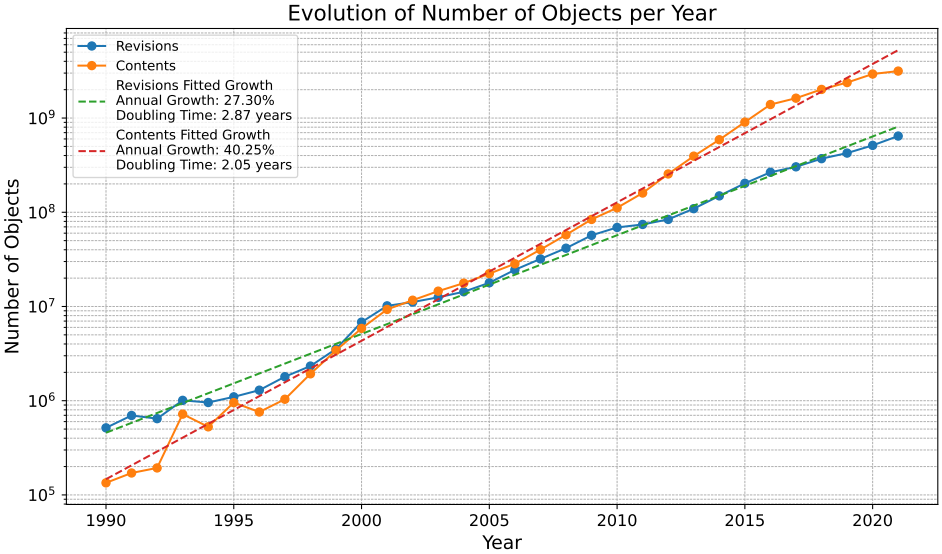
Information technology — SoftWare Hash IDentifier (SWHID) Specification V1.2

Published (Edition 1, 2025)

Criptographically strong identification for *50 billion* software artifacts

Integrity and traceability support reproducibility and transparency

Intermezzo: exponential source code growth

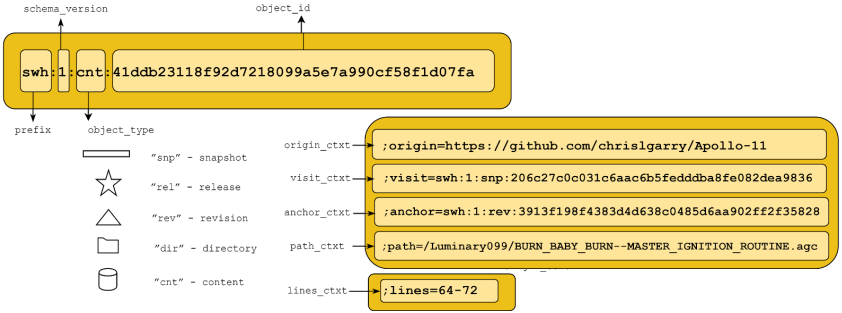


The Software Hash persistent identifier (SWHID)

Software Hash Identifiers (SWHID)

see swhid.org

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



In SPDX 2.2; IANA "swh:"; WikiData P6138; ISO standardization ongoing DIS 18670

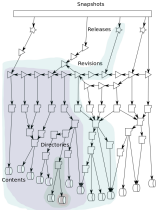
Full fledged *source code references* for traceability, integrity and reproducibility

Examples: Apollo 11 AGC, Quake III rsqrt; Guidelines available: HOWTO and ICMS 2020

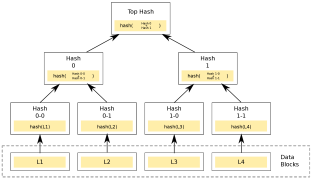
A revolutionary infrastructure

Modern "Library of Alexandria", *international, non profit, long term* initiative
addressing the needs of *industry, research, culture and society as a whole*

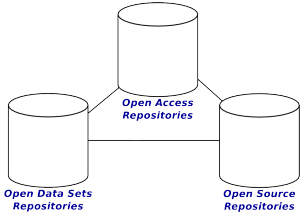
Software Graph



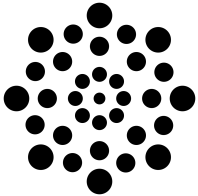
Software Blockchain



Open Science pillar



Big Code

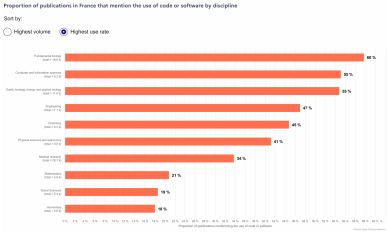


One infrastructure, *shared*: more efficient, less waste …
… addressing a broad spectrum of needs!

Outline

Software Source Code: pillar of Open Science

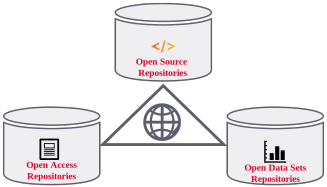
Software powers modern research



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!

Software Heritage adresses Open Science's needs

☞

Archive, Reference, Describe and Cite all (research) software!

How Software Heritage adresses Open Science's needs

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!)*

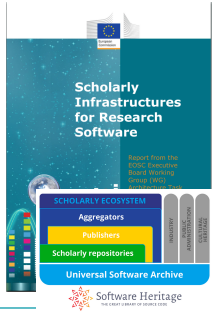
to give *credit* to authors (*evaluation*!)

Demo time!

- Browse + Reference [DIS 18670] (Apollo 11 [excerpt], your work may be already there !)
- Trigger archival, use the updateswh browser extension, configure the webhooks
- Cite from the archive with biblatex-software (CTAN, ACMART)
- 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
- Example with Parmap: devel on Github, archive in SWH, curated deposit in HAL
- Example research articles:
  - compare Fig. 1 and conclusions in the 2012 version and the updated version
  - SWHID in a replication experiment

A few adoption indicators

Policy



- Recommendations in ANR 2023 guidelines (p. 17)
- HAL+SWH in the Open Science software booklet

Projects

eosc | FAIRCORE4EOSC
Core Components Supporting a FAIR EOSC

The CodeMeta Project

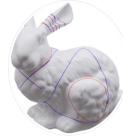
eosc | FAIR-IMPACT
Expanding FAIR solutions across EOSC

Users and collaborations

What are they "referencing"?

source	n	percentage
Not available	2868	46.22
GitHub	1151	18.55
software heritage	387	6.24
zenodo	142	2.29
r package	70	1.13
cran	56	0.90
r package version	54	0.87
gitlab	35	0.56

Graphics Replicability Stamp Initiative



b/Surf: Interactive Bézier Splines on Surface Meshes

Claudio Mancinelli, Giacomo Nazzaro, Fabio Pellacini, Enrico Puppo
IEEE Transactions on Visualization and Computer Graphics (TVCG)

doi | Repository

Outline

Analysis of the NII's contribution to Open Source

24892

Contributions

867

Software Projects

Active last 12 months: 19

Still crawled: 793

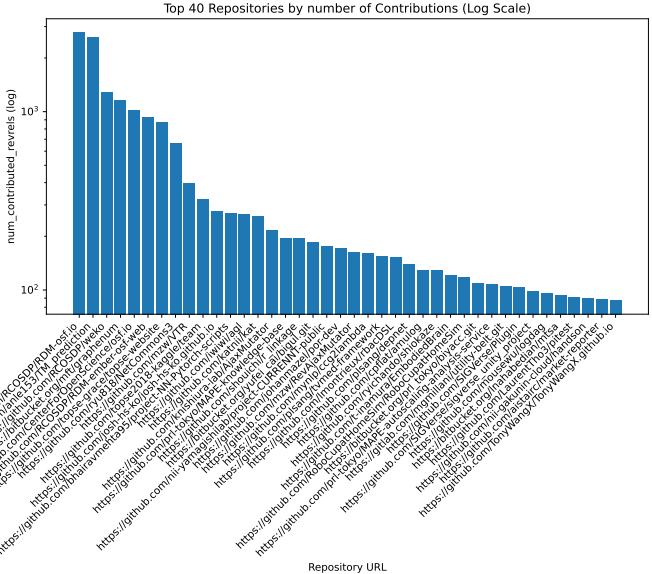
270

Contributors

2010-02-03
2024-12-01

Time Span

Analysis of the NII's contribution to Open Source, cont'd



Outline

Open source security

Open source software can be freely used, copied, and modified.

Open Source Software (OSS) is everywhere

- Huge boost for innovation! (e.g., reduced time to market)
- 96% of (non-open) software products depend on open source (2022).
- Open source is at the heart of the global digital infrastructure.

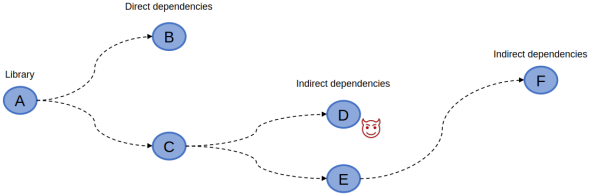
With great exposure comes great scrutiny…

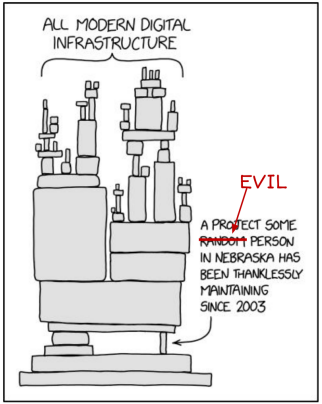
- …by both good and bad actors.
- OSS is more and more targeted by attackers.
- Increased policy attention to secure OSS, e.g.:
  - US: POTUS executive orders (2022, Jan 2025!)
  - EU: Cyber Resilience Act (published 2024, full effect 2027!)

Software supply chain attacks

Reusing OSS via dependencies

- Software dependencies: a popular way of reusing open source software.
- Software product A uses functionalities implemented in OSS product B ... and so on.





based on xkcd.com/2347

Attacking the software supply chain

- Attacking undermaintained "leaf" packages (e.g., D) → efficient attack strategy
- Many documented attacks: event-stream (2018), node-ipc (2022), XZ utils (2024), ...

Securing open source with Software Heritage

The largest archive that guarantees

1. availability,
2. integrity, and
3. traceability of (OSS) source code.

A universal, open knowledge base

- *facts* about open source software
- available to everyone (not only the big players) to secure OSS

SWHSec project (2023-2027) swhsec.github.io

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

- funded by French national CampusCyber
- 8 research teams, including SWH core
- Axes:
  - extend SWH with security information,
  - code analysis, dependency analysis,
  - vulnerability tracking … at SWH scale.

Cyber Resilience Act (policy)

Open Regulatory Compliance
Leading Regulatory Compliance for Open Source

Joined the Open Regulatory Compliance Working Group

Position paper ☞

Outline

Software Heritage and Generative AI, first contacts

October 19, 2023

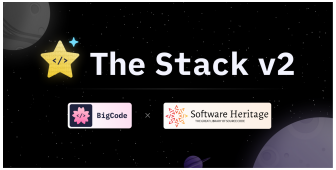
Software Heritage Statement on Large Language Models for Code

Principles

1. Knowledge derived from the Software Heritage archive must be given back to humanity, rather than monopolized for private gain. The resulting *machine learning models* must be made available under a suitable open license, together with the documentation and toolings needed to use them.
2. The *initial training data extracted from the Software Heritage archive* must be fully and precisely identified by, for example, publishing the corresponding SWHID identifiers (note that, in the context of Software Heritage, public availability of the *initial training data* is a given: anyone can obtain it from the archive). This will enable use cases such as: studying biases (fairness), verifying if a code of interest was present in the training data (transparency), and providing appropriate attribution when generated code bears resemblance to training data (credit), among others.
3. Mechanisms should be established, where possible, for authors to exclude their archived code from the training inputs before model training begins.

February 2024

Yes, it's possible!



But it's hard…

Lessons learned

Principles

1. Knowledge derived from the Software Heritage archive must be given back to humanity, rather than monopolized for private gain. The resulting *machine learning models* must be made available under a suitable open license, together with the documentation and toolings needed to use them.
2. The *initial training data extracted from the Software Heritage archive* must be clearly and precisely identified by, for example, publishing the corresponding **SWHID identifiers** (note that, in the context of Software Heritage, public availability of the *initial training data* is a given: anyone can obtain it from the archive). This will enable use cases such as: studying biases (fairness), verifying if a code of interest was present in the training data (transparency), and providing appropriate attribution when generated code bears resemblance to training data among others.
3. Mechanisms should be established, where possible, for authors to exclude their archived code from the training inputs before model training begins.

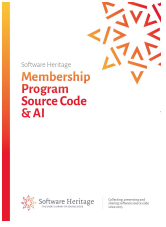
Transparency is easy: use

- SWHID (**ISO/IEC standard 18670**)
- and **Software Heritage**

→ Software Heritage SCAI ***members*** can

- **deposit** source codes
- **publish** their SWHID

today!



(Re)use of the data is tricky: who is *the real owner of a source code file*? What are the real use rights?

- **Building a qualified training set** is expensive (includes **license detection at massive scale)**
- **Attribution** on model output **is challenging**

(700 billion edges, and counting!)

→ Need **a coordinated effort**

It's time to build a

Code Commons

CodeCommons

Next generation infrastructure
for enabling transparent AI on code
and massive analysis of software source code

Inria | Software Heritage – The Great Library of Source Code | cea | TWEAG by Modus Create

Sant'Anna School of Advanced Studies – Pisa | Università degli Studi di Torino | Università di Pisa | Alma Mater Studiorum Università di Bologna | AboutCode

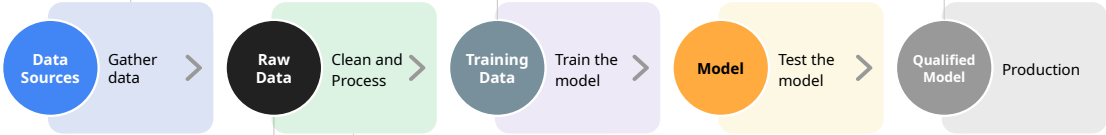
GENCI – HPC at the service of knowledge | bpifrance – Servir l'avenir | France 2030

A STEP FORWARD: CodeCommons

Availability: Easy access to all relevant data for software (source code, PR, issues, discussions, etc.) **Shared harvesting:** we do it only once!

Traceability: Identify and make available the data used for training

Ethics: Provide tools to verify the provenance and attribution of generative AI outputs

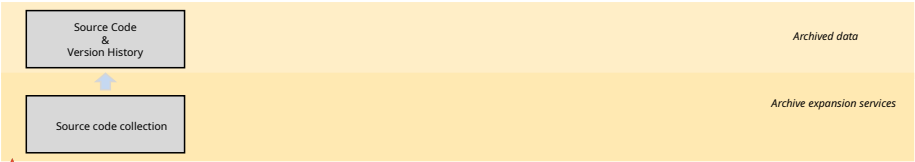


Structuring: Organize and connect the various data sources to create a coherent training set.

Efficiency: Facilitate the extraction of qualified datasets to build high-performance models.

Solutions

CodeCommons: bird’s eye view (technical focus)

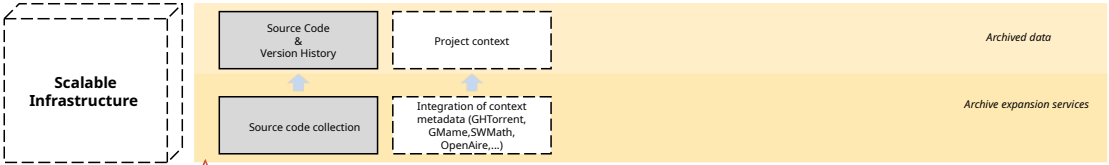


Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

CodeCommons: bird’s eye view (technical focus)

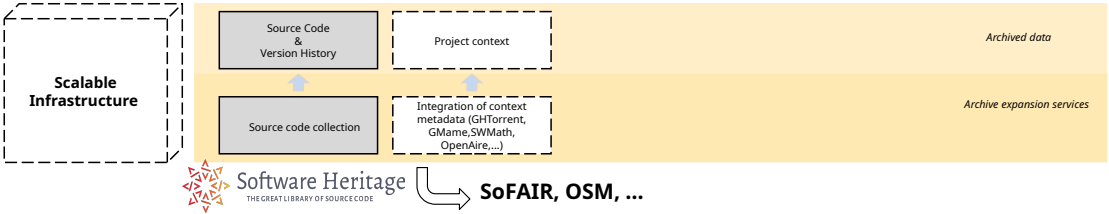
Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

CodeCommons: bird’s eye view (technical focus)

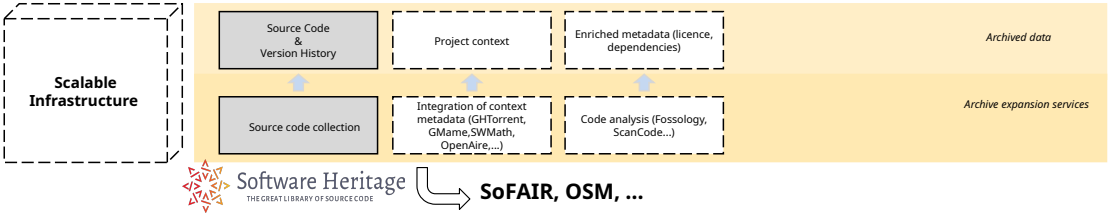


Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

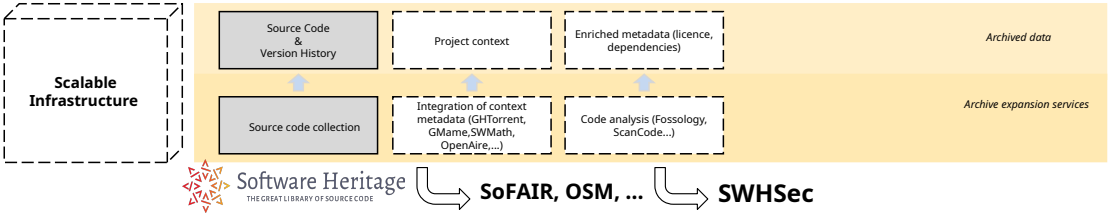
CodeCommons: bird’s eye view (technical focus)



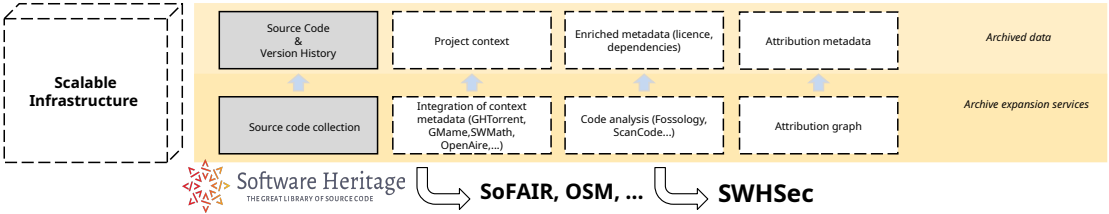
CodeCommons: bird’s eye view (technical focus)



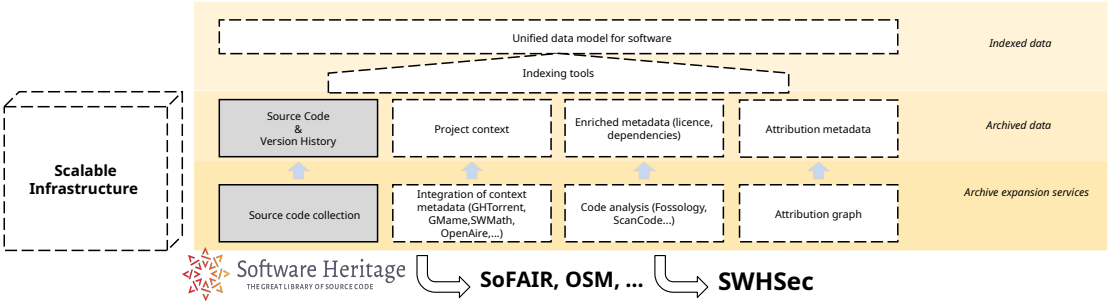
CodeCommons: bird’s eye view (technical focus)



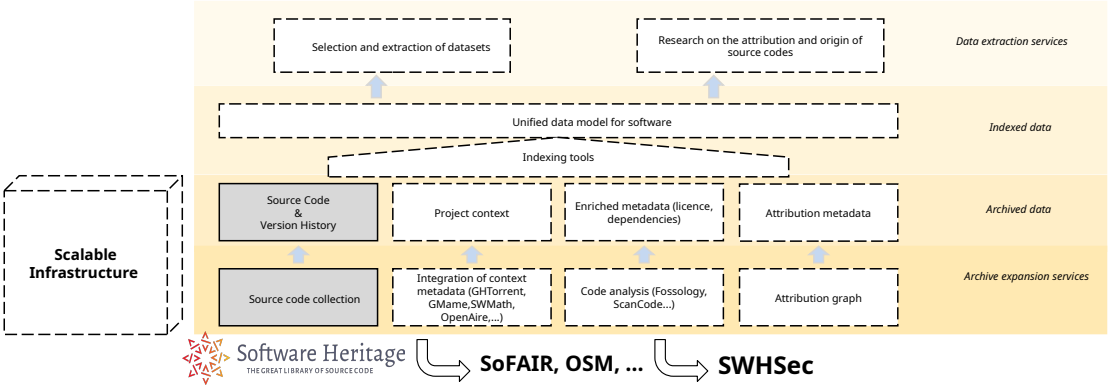
CodeCommons: bird’s eye view (technical focus)



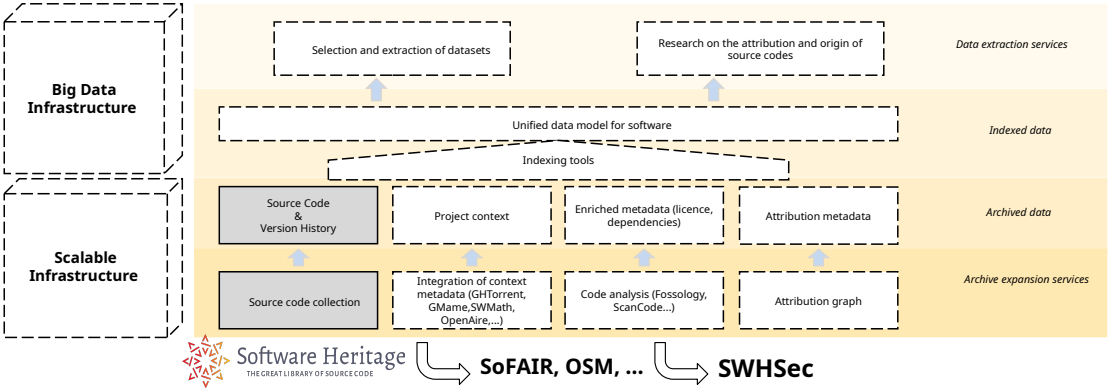
CodeCommons: bird’s eye view (technical focus)



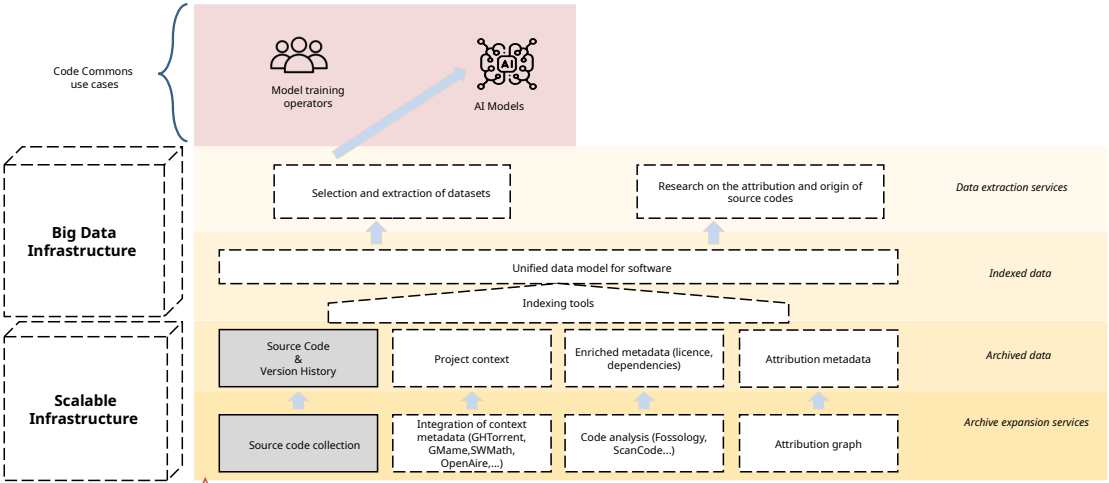
CodeCommons: bird’s eye view (technical focus)



CodeCommons: bird’s eye view (technical focus)



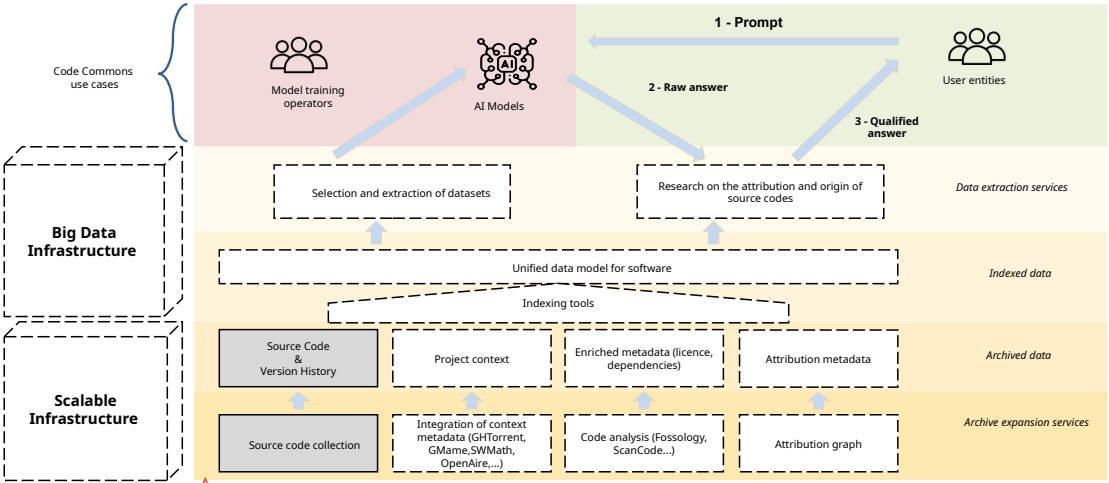
CodeCommons: bird’s eye view (technical focus)



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

SoFAIR, OSM, ... → **SWHSec**

CodeCommons: bird’s eye view (technical focus)



CodeCommons

Open, responsible, and transparent AI: Our shared goal

CodeCommons is an ambitious project to create the world's most comprehensive digital commons for code

Building on the existing foundation of Software Heritage, the largest publicly available source code archive, CodeCommons aims to bring into one place all the **critical** and **qualified** information needed to create **smaller**, **better** datasets for the next generation of AI tools.

At its core, the project prioritizes transparency and traceability, enabling model builders and users to **respect creators' rights** while promoting **sovereign** and **sustainable** AI.

Learn more

Meet the teams

Outline

Chile launches national AI policy and introduces AI bill following UNESCO´s recommendations.

Chile stands at the forefront of global artificial intelligence (AI) governance efforts, with UNESCO playing a pivotal role in shaping and advancing its initiatives.

UNESCO

4 May 2024 - Last update:6 May 2024

Fostering Open Science and Responsible AI: A Week of High-Level Dialogues and New Alliances in Chile

Kenneth Pugh • 1st 1mo
Senador en Senado de la República de Chile

Thanks **Roberto Di Cosmo** for the **#SWHeritage** project and the endless uses for **#ResponsibleAI** to achieve **#DigitalSoveranaty**.

Choose Europe for Science

Sorbonne, Paris, May 5th 2025

Address of President Macron

ÉLYSÉE

En matière de climat, de connaissances de la planète, en matière de sciences du vivant, nous avons besoin de poursuivre ce travail, de consolider les plateformes et d'avancer. Réfléchissons collectivement à la lourde tâche d'archivage de certains savoirs spécifiques. Je pense aux travaux aussi de *Software Heritage*, qui, en France, archive l'ensemble des logiciels créés dans le monde, avec une initiative européenne à construire sur le sujet. Je souhaite, au-delà de la question de notre capacité, à attirer chercheurs, enseignants-chercheurs, que nous ayons aussi une capacité à préserver, consolider, attirer plateformes de collaboration et données au service de l'intérêt général.

A unique opportunity!

France - Japan collaboration opportunities

Software Pillar of Open Science

- *JairoCloud and Software Heritage*

Large Language Models

- Engage with *Code Commons*

Access all public source code

- Long term *reproducibility* for science
  - *NII RDC and Software Heritage*
- *Resilience* to code hosting outage

Call to action

- include in Open Science policy
- endorse for transparent AI
- leverage for CyberSecurity
- join the members community

Annual report 2024

5 years in 5 minutes

A growing and active community

Core Team

All together, 2025 Symposium at UNESCO

Ambassadors

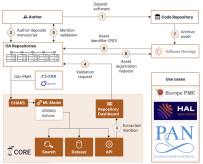
Becoming an Ambassador

Interested in becoming a Software Heritage ambassador?
Tell us about yourself and your interest in our mission.

ambassadorprogram@softwareheritage.org

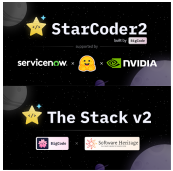
Software Heritage: 2024 progress hilights

January 2024 — SoFAIR Kickoff



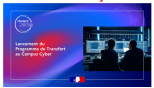
Massive identification, archival and reference of software in academic publications

February 2024 — First AI contact



StarCoder2 trained on GitHub subset of Software Heritage

February 2024 — SWHSec Kickoff



SWHSec: 5 years PTCC project

Software supply chain Security

Academic interoperability

Connect with Zenodo, Episcience, Dagsthul, SwMath, OpenAire, …

Fall 2024 — CodeCommons

Building the reference software knowledge base for AI

… and much more

2024 annual report here →

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

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

Software Heritage enables

- archival, reference, integrity
- qualification, sharing and reuse

Call to action

- include in Open Science policy
- endorse for transparent AI
- leverage for CyberSecurity
- join the members community

Annual report 2024

5 years in 5 minutes