Software Heritage

revolutionary infrastructure for our digital future

Roberto Di Cosmo

Director, Software Heritage
Inria and Université Paris Cité

January 29 2025
UNESCO

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

Software is built from *Source Code*

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."

Software is built from *Source Code*

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
```

Software is built from *Source Code*

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;
};
```

Software is built from *Source Code*

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

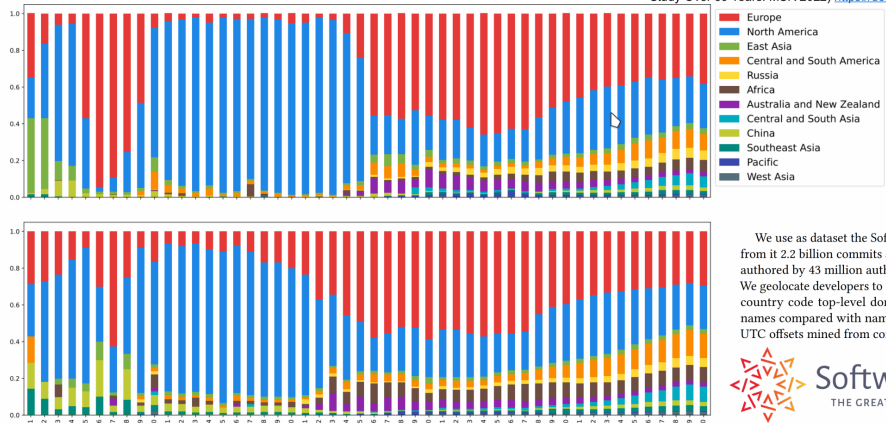
"Source code provides a view into the mind of the designer."

(Open) Source Code comes from all over the world…

MSR '22, May 23–24, 2022, Pittsburgh, PA, USA

Davide Rossi and Stefano Zacchiroli

Geographic Diversity in Public Code Contributions: An Exploratory Large-Scale Study Over 50 Years. MSR 2022) https://doi.org/10.1145/3524842.3528471

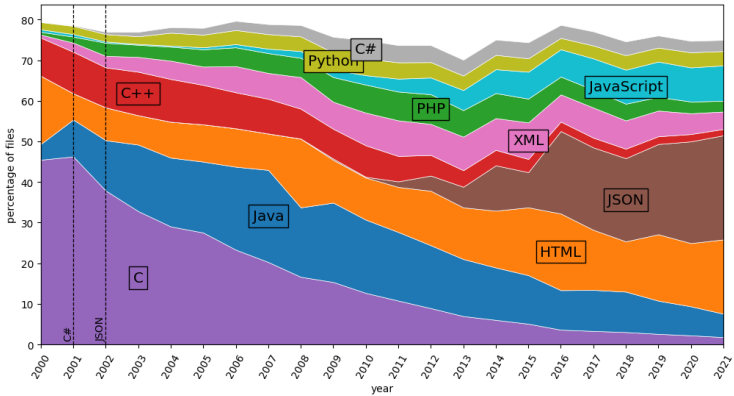


We use as dataset the Software Heritage archive [3] and analyze from it 2.2 billion commits archived from 160 million projects and authored by 43 million authors during the 1971–2021 time period. We geolocate developers to 12 world regions, using as signals email country code top-level domains (ccTLDs) and author (first/last) names compared with name distributions around the world, and UTC offsets mined from commit metadata.

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Figure 3: Ratio of commits (above) and active authors (below) by world zone over the 1971–2020 period.

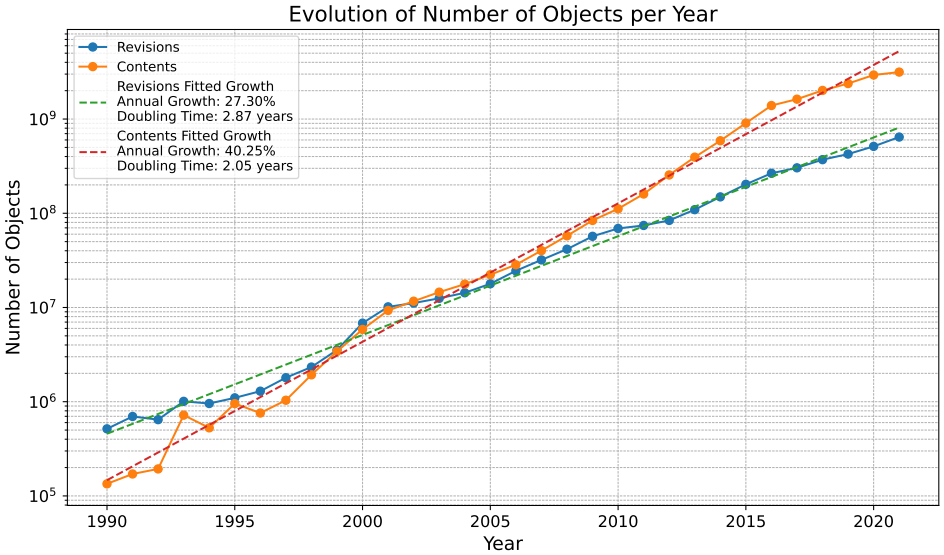
… it is written in many (programming) languages …



Evolution of the activity for programming, markup, and data languages from 2000 to 2021.

A. Desmazières, R. Di Cosmo, V. Lorentz
50 years of programming language evolution through the Software Heritage Looking Glass
MSR 2025. To appear.

… and it grows at an exponential rate



(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

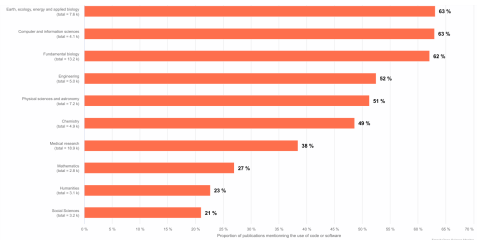
"*The real antidote [to epidemic] is* scientific knowledge *and* global cooperation."

(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

"*The real antidote [to epidemic] is* scientific knowledge *and* global cooperation."

Software powers modern research



20%+ articles use software, all disciplines

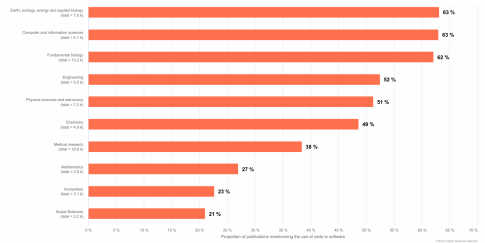
2023 French Open Science Monitor

(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

"The real antidote [to epidemic] is scientific knowledge *and* global cooperation."

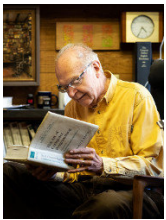
Software powers modern research



20%+ articles use software, all disciplines

2023 French Open Science Monitor

We can still talk to the early inventors



"Telling historical stories is the best way to teach. It's much easier to understand something if you know the threads it is connected to."

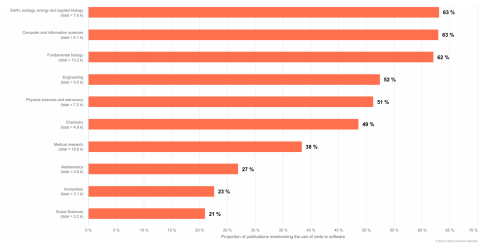
Donald E. Knuth
Len Shustek
CACM, January 2021

(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

"*The real antidote [to epidemic] is* scientific knowledge *and* global cooperation."

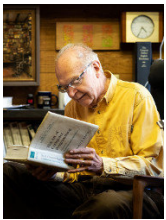
Software powers modern research



20%+ articles use software, all disciplines

2023 French Open Science Monitor

We can still talk to the early inventors



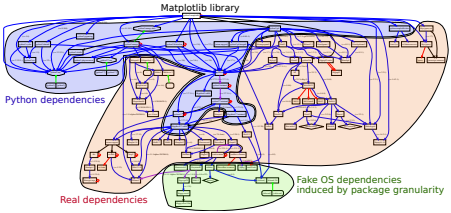
"Telling historical stories is the best way to teach. It's much easier to understand something if you know the threads it is connected to."

Donald E. Knuth
Len Shustek
CACM, January 2021

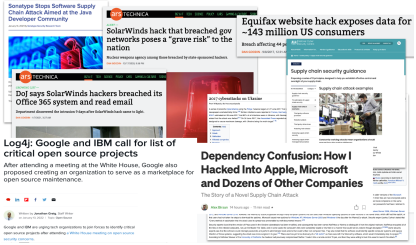
We need a *dedicated infrastructure* to preserve and share *all* this knowledge!

Enhancing software Reuse, Security and Transparency

Software complexity is growing…

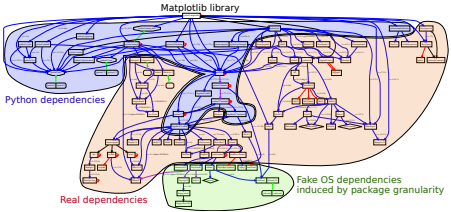


…it is important to Know Your SoftWare (KYSW)



Enhancing software Reuse, Security and Transparency

Software complexity is growing…



…it is important to Know Your SoftWare (KYSW)

Sec. 4. Enhancing Software Supply Chain Security

ensuring and attesting […] to the integrity and provenance of open source software

May 2021 POTUS Executive Order

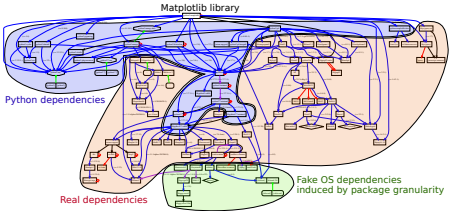
Cyber Resilience Act

Regulation aims to […] ensuring […] software products […] with fewer vulnerabilities.

REGULATION (EU) 2024/2847

Enhancing software Reuse, Security and Transparency

Software complexity is growing…



…it is important to Know Your SoftWare (KYSW)

Sec. 4. Enhancing Software Supply Chain Security

ensuring and attesting […] to the integrity and provenance of open source software

May 2021 POTUS Executive Order

Cyber Resilience Act

Regulation aims to […] ensuring […] software products […] with fewer vulnerabilities.

REGULATION (EU) 2024/2847

We need a *trusted* knowledge base with *software integrity and provenance*!

Software source code is fragile

Endangered source code …

- *link rot*
- *data rot*
- *platform consolidation*
  - 2015 Google Code and Gitorious.org shutdown: ~1M
  - 2019 Bitbucket mercurial phase out: ~250.000
  - 2022 GitLab.com: remove inactive projects?

Software source code is fragile

Endangered source code …

- *link rot*
- *data rot*
- *platform consolidation*
  - 2015 Google Code and Gitorious.org shutdown: ~1M
  - 2019 Bitbucket mercurial phase out: ~250.000
  - 2022 GitLab.com: remove inactive projects?

… is endangered knowledge!

broken links and missing pieces in the *web of knowledge* of humankind

Software source code is fragile

Endangered source code …

- *link rot*
- *data rot*
- *platform consolidation*
  - 2015 Google Code and Gitorious.org shutdown: ~1M
  - 2019 Bitbucket mercurial phase out: ~250.000
  - 2022 GitLab.com: remove inactive projects?

… is endangered knowledge!

broken links and missing pieces in the *web of knowledge* of humankind

Bottomline: we need a global, long term effort

to build a *universal archive* of *all software source code*
make it *resilient*
and make it *sustainable*

Outline

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

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

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

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

Today: a *universal* software archive, as a shared infrastructure

One infrastructure
open and shared

Inria

unesco

Today: a *universal* software archive, as a shared infrastructure

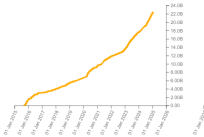
One infrastructure
open and shared

Inria

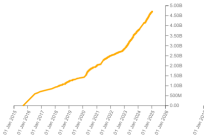
unesco

The largest archive ever built

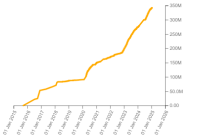
Source files
22,548,654,226



Commits
4,743,656,085



Projects
344,672,354



Directories
17,848,569,305

Authors
86,143,084

Releases
102,056,508

Today: a *universal* software archive, as a shared infrastructure

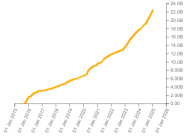
One infrastructure
open and shared

Inria

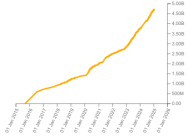
unesco

The largest archive ever built

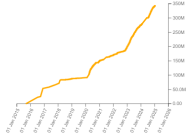
Source files
22,548,654,226



Commits
4,743,656,085



Projects
344,672,354



Directories
17,848,569,305

Authors
86,143,084

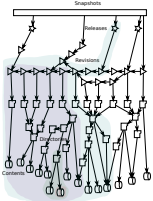
Releases
102,056,508

Bitbucket 2,578,475 origins	 56,975 origins	git 33,350 origins
R 27,377 origins	debian 141,834 origins	 88,561 origins
GitHub 249,868,189 origins	gitiles 24,378 origins	GitLab 5,736,223 origins
git 3,791 origins	Gogs 394 origins	GO 1,887,337 origins
Guix 68,391 origins	GNU 354 origins	heptapod 1,340 origins
launchpad 654,755 origins	Maven 312,179 origins	NixOS 48,905 origins
npm 4,003,267 origins	 5,438 origins	Packagist 376,882 origins

figures as of January 18 2025

A revolutionary infrastructure

The *graph* of public software development

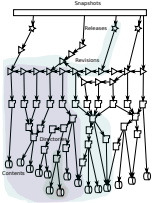


All software development
in a single graph …

- enable traceability

A revolutionary infrastructure

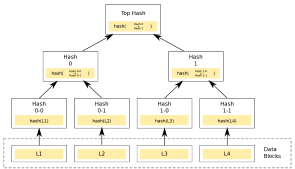
The *graph* of public software development



All software development in a single graph …

- enable traceability

The *global ledger* of public code

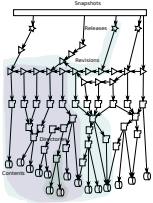


… a Merkle graph

- ensure integrity

A revolutionary infrastructure

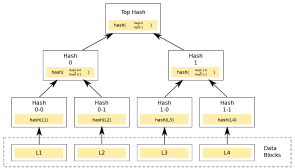
The *graph* of public software development



All software development in a single graph …

- enable traceability

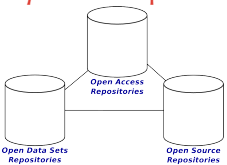
The *global ledger* of public code



… a Merkle graph

- ensure integrity

A *pillar* of Open Science

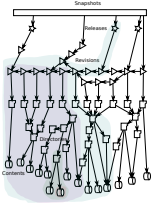


Reference archive of Research Software

- reproducibility
- reference

A revolutionary infrastructure

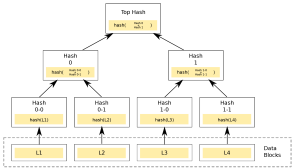
The *graph* of public software development



All software development in a single graph …

- enable traceability

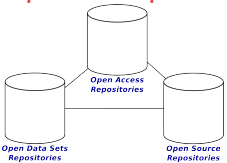
The *global ledger* of public code



… a Merkle graph

- ensure integrity

A *pillar* of Open Science

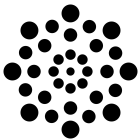


Reference archive of Research Software

- reproducibility
- reference

Reference platform for *Big Code*

uniform data structure



- large scale studies
- cybersecurity
- machine learning, AI, …

more later today

Sharing the vision

And many more ...

www.softwareheritage.org/support/testimonials

Sharing the vision

United Nations Educational, Scientific and Cultural Organization

Donors, members, sponsors

Diamond sponsors

Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors

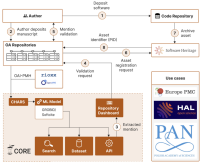
And many more ...

www.softwareheritage.org/support/testimonials

we are all concerned, anyone can join and help

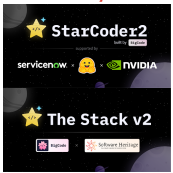
Software Heritage: 2024 progress hilights

January 2024 — SoFAIR Kickoff



Massive identification, archival and reference of software in academic publications

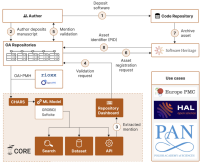
February 2024 — First AI contact



StarCoder2 trained on GitHub subset of Software Heritage

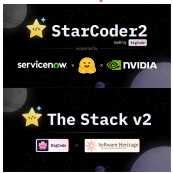
Software Heritage: 2024 progress hilights

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

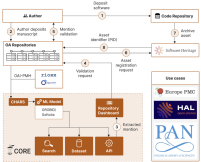
Academic interoperability

eosc
FAIR-IMPACT
FAIRCORE4EOSC

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

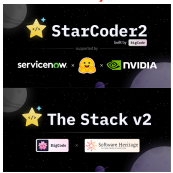
Software Heritage: 2024 progress hilights

January 2024 — SoFAIR Kickoff



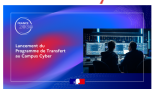
Massive identification, archival and reference of software in academic publications

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

EOSC
FAIR-IMPACT
FAIRCORE4EOSC

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 growing and active community

Core Team

All together, 2024 Symposium

Ambassadors

Outline

Cybersecurity and regulations

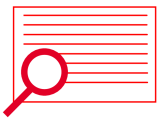
- Securing Critical Software
- Cyber Resilience Act
- Preserving Open Source
- Research and Tools

Today's main topics from the connecting hub

Cybersecurity and regulations

- Securing Critical Software
- Cyber Resilience Act
- Preserving Open Source
- Research and Tools

AI, transparency and regulations



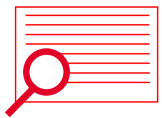
- Open Source AI
- AI Act
- Data transparency
- The global view

Today's main topics from the connecting hub

Cybersecurity and regulations

- Securing Critical Software
- Cyber Resilience Act
- Preserving Open Source
- Research and Tools

AI, transparency and regulations



- Open Source AI
- AI Act
- Data transparency
- The global view

Open Science

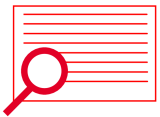
- Policy and implementation
- Monitoring and recognizing
- Computational reproducibility

Today's main topics from the connecting hub

Cybersecurity and regulations

- Securing Critical Software
- Cyber Resilience Act
- Preserving Open Source
- Research and Tools

AI, transparency and regulations



- Open Source AI
- AI Act
- Data transparency
- The global view

Open Science

- Policy and implementation
- Monitoring and recognizing
- Computational reproducibility

Memory of the World

- Preserving software history
- Software to preserve history
- The role of AI

Outline

Outline

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

Software Heritage is the foundation…

- vendor neutral, multi-stakeholder
- open source, non profit
- a worldwide initiative
- a long term initiative

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

Software Heritage is the foundation…

- vendor neutral, multi-stakeholder
- open source, non profit
- a worldwide initiative
- a long term initiative

… that enables

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

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

Software Heritage is the foundation…

- vendor neutral, multi-stakeholder
- open source, non profit
- a worldwide initiative
- a long term initiative

… that enables

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

The way forward

Scale up Software Heritage to serve all stakeholders worldwide

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

Software Heritage is the foundation…

- vendor neutral, multi-stakeholder
- open source, non profit
- a worldwide initiative
- a long term initiative

… that enables

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

The way forward

Scale up Software Heritage to serve all stakeholders worldwide

You can help!

use, adopt, advocate, contribute, fund, support, join

A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

Software Heritage

www.softwareheritage.org @swheritage

The Library of Alexandria of code

- recover the past
- structure the future
- rebuild trust in science

The Very Large Telescope for Source code

- explore and reuse
- better, more secure software

for society as a whole

Inria Inria La Fondation unesco