Software Heritage

Sponsor and Members General Assembly

Roberto Di Cosmo
Director
roberto@dicosmo.org

February 18th, 2026

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

Software is all around us

Invisible fabric of digital society

Knowledge is in the source

```
/**
 * @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=
           Global search and replace of 'unsigned short int' with 'int' would remove this limit, but use more memo
        */

        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) with
        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;
};
```

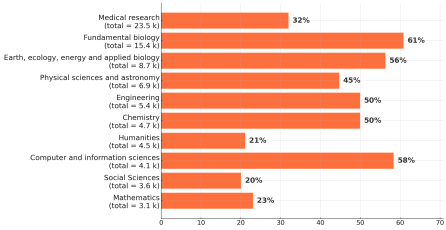
Covid Sim (excerpt)

Len Shustek, Computer History Museum 2006

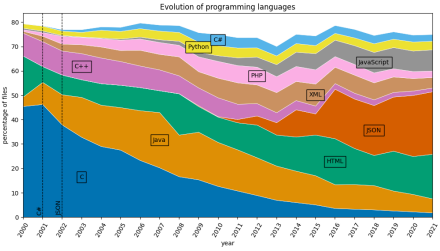
“Source code provides a view into the mind of the designer.”

A Global Undertaking

Pillar of Science across all research areas



Many programming languages



From all continents

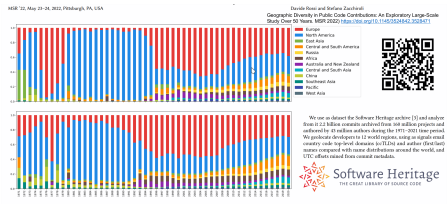
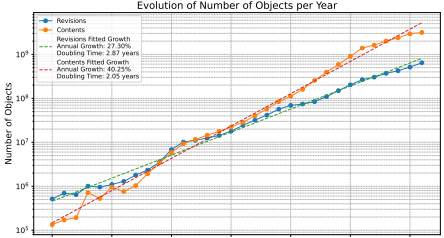


Figure 3: Ratio of commits (above) and active authors (below) by world zone over the 1971–2020 period.

An exponential growth



(open source) software and research

Nature, October 2025

nature

View all journals | Search | Log in

Explore content | About the journal | Publish with us | Subscribe | Sign up for alerts | RSS feed

nature › comment › article

COMMENT | 07 October 2025

Stop treating code like an afterthought: record, share and value it

Scientists, research institutions, funders, libraries and publishers must all improve software practices.

By Roberto Di Cosmo, Sabrina Granger, Konrad Hinsen, Nicolas Jullien, Daniel Le Berre, Violaine Louvet, Camille Maumet, Clémentine Maurice, Raphaël Monat & Nicolas P. Rougier

Illustration: Phil Wheeler

ICSE FoSE 2026

Reclaiming Software Engineering as the Enabling Technology for the Digital Age

Tanja E. J. Vos (1) , Tijs van der Storm (2) , Alexander Serebrenik (3) , Lionel Briand (4) , Roberto Di Cosmo (5) , J.-M Bruel (6) , Benoît Combemale (7, 6)

Show details

1. Open University of the Netherlands [Heerlen]
2. CWI - Centrum Wiskunde & Informatica
3. TU/e - Eindhoven University of Technology [Eindhoven]
4. LERO - The Irish Software Engineering Research Centre
5. Software Heritage
6. IRIT-SM@RT - Smart Modeling for softw@re Research and Technology
7. DiverSe - Diversity-centric Software Engineering

Abstract en

Software engineering is the invisible infrastructure of the digital age. Every breakthrough in artificial intelligence, quantum computing, photonics, and cybersecurity relies on advances in software engineering, yet the field is too often treated as a supportive digital component rather than as a strategic, enabling discipline. In policy frameworks, including major European programmes, software appears primarily as a building block within other technologies, while the scientific discipline of software engineering remains largely absent. This position paper argues that the long-term sustainability, dependability, and sovereignty of digital technologies depend on investment in software engineering research. It is a call

Keywords en

open source | sustainability | digital sovereignty | research policy | enabling technology | software engineering

Domains

Software Engineering [cs.SE]

Complete list of metadata

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

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

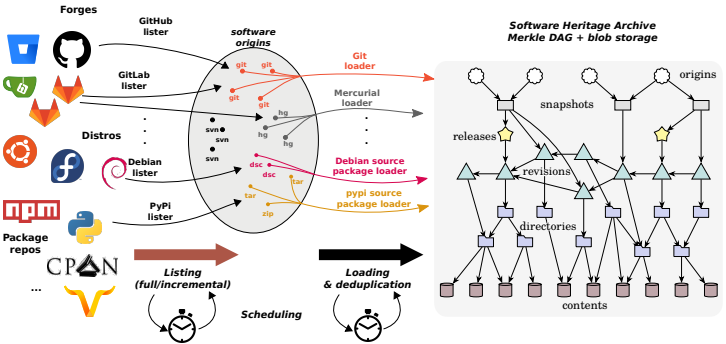
Inria unesco

The largest archive ever built

Bitbucket 2,979,416 origins	56,983 origins	git 33,507 origins
30,750 origins	debian 145,690 origins	101,132 origins
GitHub 300,899,172 origins	gitiles 25,242 origins	GitLab 5,906,407 origins
git 3,973 origins	Gogs 494 origins	GO 2,326,857 origins
Guix 75,284 origins	GNU 354 origins	heptapod 1,391 origins
launchpad 664,326 origins	Maven 520,999 origins	NixOS 73,325 origins
npm 4,719,410 origins	5,175 origins	Packagist 384,120 origins
Redox PAGURE 72,459 origins	Phabricator 198 origins	pub.dev 73,901 origins

figures as of January 8 2026

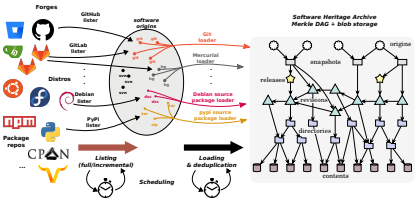
From a Babel tower to a giant graph



Global development history permanently archived in a uniform data model

- over 27 billion unique source files from over 420 million software projects
- ~2PB (compressed) blobs, ~50 B nodes, ~1 trillion edges

Archive (25B+ files, 400M+ projects)



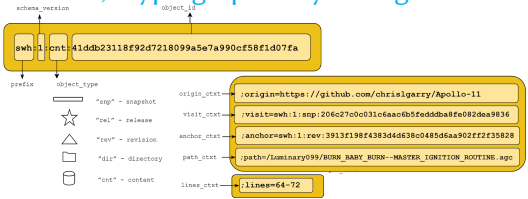
- save now, updateswh, webhooks
- deposit.softwareheritage.org

Describe

- *Intrinsic metadata* from source code
- Contributed the Codemeta generator

Reference (50 billion SWHIDs)

Intrinsic, cryptographically strong IDs



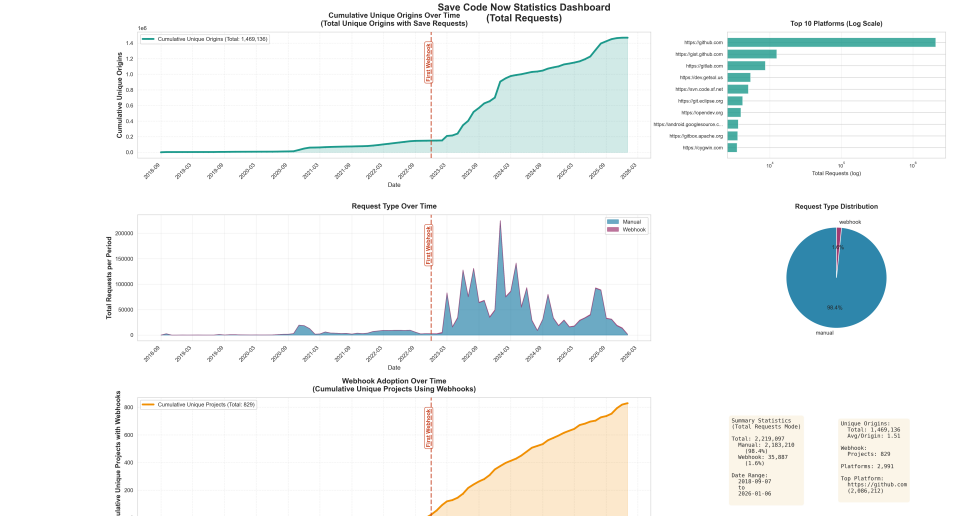
Now in SPDX 2.2, Wikidata

`https://swhid.org` - ISO/IEC 18670

Cite/Credit

- Contributed biblatex-software style
- Software Citation from the archive!

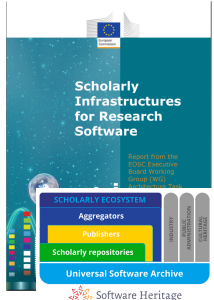
A peek at Save Code Now adoption



Outline

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

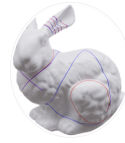
eosc | FAIR-IMPACT
Expanding FAIR solutions across EOSC

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

Upcoming products and services

Institutional portal

OSPO RADAR

Vulnerability graph

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Strategic Insights

Software Heritage

Strategic Software Insights Report

Dataset Factory

</> </>
</> </> </>
</> </>

See annual report for more ☞

Outline

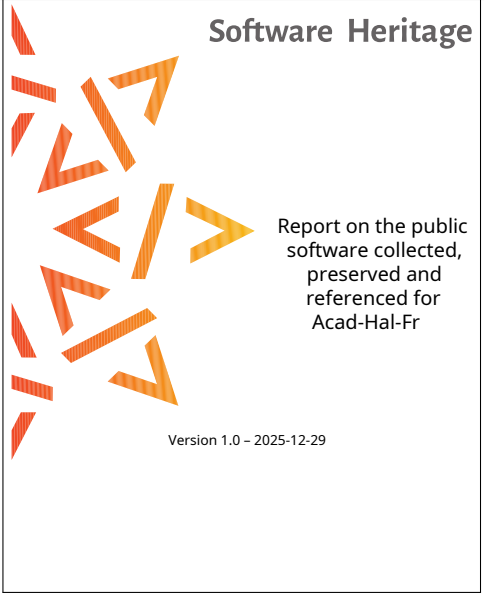
Strategic Software Insights report: why it matters?

Software is a strategic asset, but

- … tracking and analysing your software assets is difficult
- … assessing your open source presence is hard
- … knowing how you compare to other is complex

Software Heritage helps you gain deep insight into your software assets

A preview of the Strategic Software Insights report



Contents

A peek inside the French Academic Software report

3531626

Contributions

46533

Software Projects

Active last 12 months: 3425

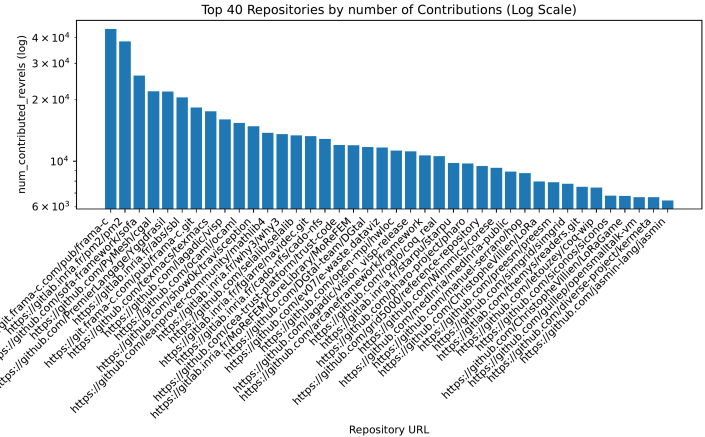
Still crawled: 44874

30101

Contributors

1992-07-08
2025-05-17

Time Span

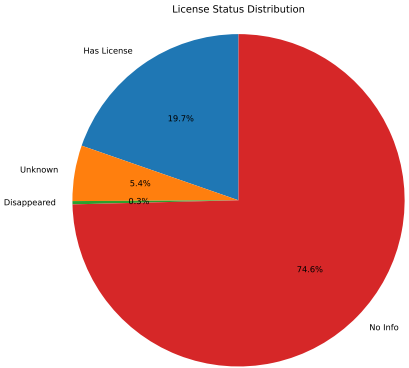


domain	count
inria.fr	3833
univ-lyon1.fr	2806
univ-orleans.fr	1555
uca.fr	1371
jussieu.fr	1247
utc.fr	1216
cea.fr	1187
insa-toulouse.fr	747
ens-lyon.fr	706
univ-artois.fr	696
inrae.fr	690
univ-lemans.fr	688
cnrs.fr	682
ens.fr	675
inra.fr	644
univ-fcomte.fr	634
unilim.fr	541
insa-rennes.fr	491
univ-avignon.fr	489
univ-pau.fr	483
utt.fr	472
polytechnique.fr	437
uha.fr	422
pasteur.fr	420
telecom-paris.fr	412
univ-cotedazur.fr	363
univ-smb.fr	352
univ-tours.fr	350
u-picardie.fr	349
enpc.fr	328
sciencespo.fr	327
universite-paris-saclay.fr	323
univ-poitiers.fr	312
u-paris.fr	289
univ-angers.fr	267
univ-paris8.fr	264
insarm.fr	211
centralesupelec.fr	192
ird.fr	171
imt-atlantique.fr	159
onera.fr	158
univ-tlse2.fr	131
irstea.fr	130
enac.fr	126
univ-eiffel.fr	125
agroparistech.fr	105
cnam.fr	104
mnhn.fr	95
uphf.fr	94
univ-fcomte.fr	89

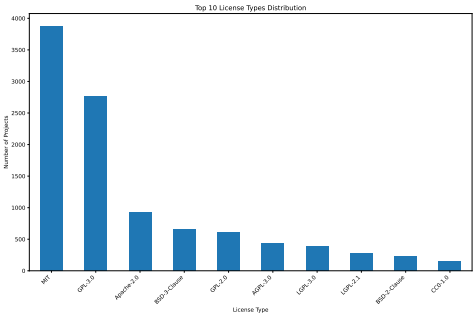
Note: Showing top 50 of 53 total domains

A peek inside the French Academic Software report, cont'd

How many projects have a license?

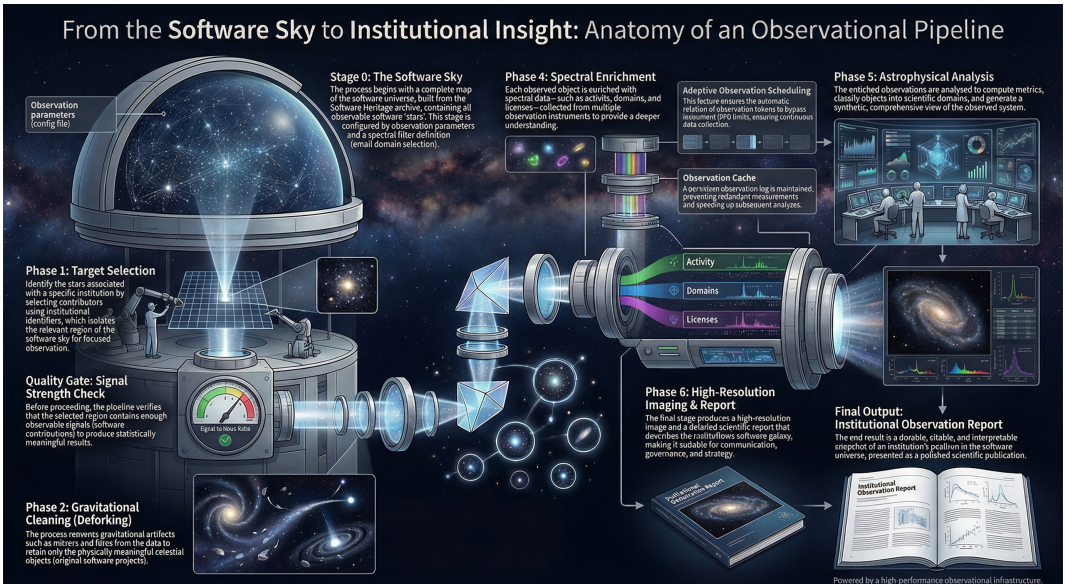


What are the top licences used?



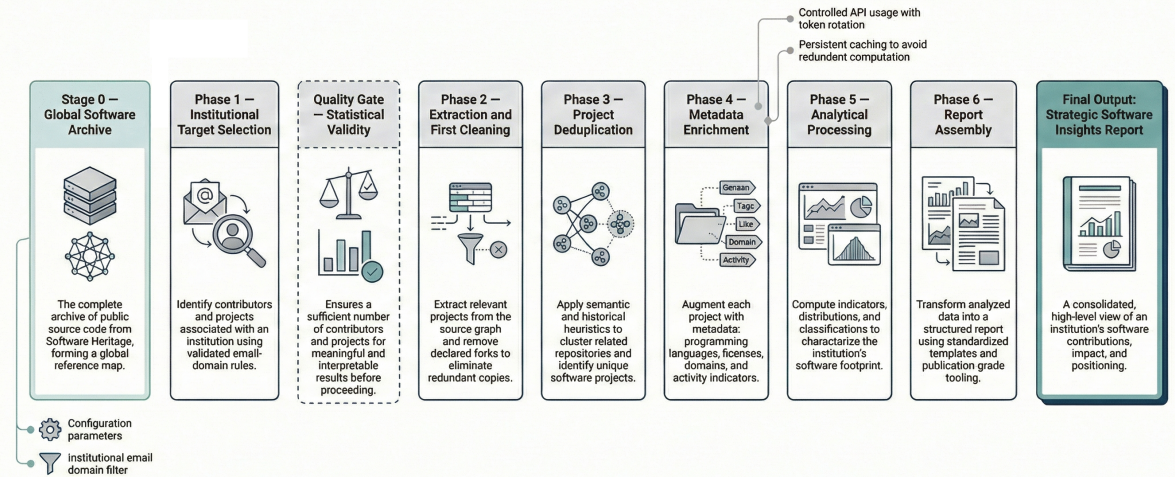
how do we build these indicators?

Software Heritage is the Very Large Telescope for source code



Strategic Software Insights Report — From Global Archive to Institutional Insight

A multi-stage analytical pipeline built on the Software Heritage archive



A complex, costly process scans the whole Software Heritage archive to deliver key insights

The report in context, and your input welcome

CodeCommons (BPI project, Datasets for AI and beyond)

- broad spectrum of metadata: licence, language, vulnerabilities, issues, PR…
- connections with context:
  - forge metadata: GitHub, GitLab, etc.
  - curated metadata: HAL, OSPO-Radar, catalogue, etc.
  - qualified software mentions in publications: Hal, swMath, SoFAIR, OpenAlex, etc.
  - vulnerabilities

Enriched Data Model

Question: what extra (meta)data is needed/useful for you?

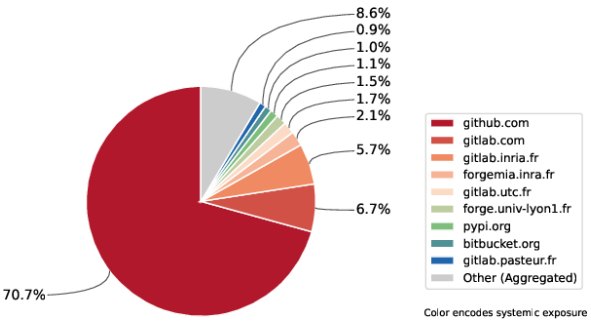
Report generation constraints

- first phase = filtering 100+M personal emails (cannot open this)
- Software Heritage must do the heavy lifting for you

Important: it will be a professional service model, with a cost

The next frontier: from preservation and analysis to resilience

Infrastructure concentration (French academia)



Collaborative development depends massively
on *very few* ~~*single*~~ *points of failures*

Software Heritage is the systemic response

Today

- All the code, also niche and legacy code
- ISO 18670 for robust traceability
- Enables long term reproducibility

With significant more funding

- automatic recovery (registry failure, network partition)
- resilience (mirrors)
- strategic autonomy