For 10 years, we have built the foundational library of all open-source code, securing humanity's shared digital knowledge.

As we look ahead, the Software Heritage Archive stands as a strategic pillar for digital sovereignty in Europe, fostering innovation and resilience for the next decade and beyond.

Director's letter

© Inria / Photo B. Fourrier

10 years of Software Heritage

In 2014, Stefano Zacchiroli and I saw a digital dark age coming: the entire history of public code was unsecured. The web got preservation efforts, but the source code — the engine of modern life — was overlooked. That crisis sparked a mission. We had to build the definitive Archive, contrasting decades of neglect with a preservation effort of radical scale.

As a computer scientist who has spent his career building, studying, and defending free and open-source software, I have seen too often what happens when digital knowledge vanishes — platforms shut down, servers disappear, maintainers move on.

With the early and decisive support of Inria, we launched Software Heritage in June 2016 and established a key partnership with UNESCO in April 2017 with an ambitious mission: to collect, preserve, and share all software source code, for the long term and for the benefit of everyone.

Building a universal commons

The 10 years that followed have been an extraordinary journey. The Archive has grown into the largest collection of source code ever assembled, now preserving more than 56 billion unique source files and over 415 million software projects — a scale unimaginable in 2016. What began as a scientific vision has evolved into a global infrastructure used across research, industry, public administration, education, and culture.

The team, which has grown from two people in 2014 to around 20 today, drives this expansion through tireless work and the steadfast support of our partners. Inria continues to provide a world-class research environment. UNESCO recently renewed its partnership, reaffirming Software Heritage's crucial role in preserving humankind's digital knowledge. This multi-stakeholder spirit is central to our mission, sustained by an expanding network of sponsors, ambassadors, researchers, and community contributors.

We've also strengthened our leadership team: Stefano is now Chief Scientific Officer, guiding long term research; Morane Gruenpeter directs scholarly ecosystems to support Open Science; David Douard oversees the growing mirror network; Thomas Aynaud joins as Chief Technology Officer to energize platform engineering; and Bastien Guerry, former French Chief Free Software Officer, will develop strategic partnerships and services for industry and public administration.

The Archive's role in ensuring continuity also became even more essential in today's uncertain geopolitical and technological landscape. Around the world, critical scientific datasets, registries, and code-hosting platforms face risks ranging from policy changes to commercial closures.

Establishing trust: The SWHID standard

Preserving source code is meaningless without ensuring its integrity and traceability. In 2025, a critical objective was met: the SoftWare Hash IDentifier (SWHID) was successfully established as the ISO/IEC 18670 international open standard for identifying all software artifacts.

This achievement validates years of intense collaboration and solidifies the SWHID as a central pillar of global digital infrastructure. Already adopted by publishers, research platforms, and AI initiatives, the identifier is quickly becoming the universal method for referencing code.

The Archive's role in ensuring continuity also became even more essential in today's uncertain geopolitical and technological landscape. Around the world, critical scientific datasets, registries, and code-hosting platforms face risks ranging from policy changes to commercial closures. Europe, in particular, has recognized the strategic importance of preserving and securing access to public code. Initiatives such as "Choose Europe for Science," the EU Cyber Resilience Act, NIS2, and the Digital Compass 2030 point to a clear need: independent, transparent, and resilient infrastructures for software.

Software Heritage stands uniquely positioned to address this challenge.

Our network of international mirrors — launched with ENEA in Italy and expanding with new partners — creates a resilient, distributed preservation fabric. Our SWHID infrastructure provides verifiable integrity for billions of components. And our collaboration with public authorities and industry groups supports compliance, reproducibility, and trust in the software supply chain.

The future of responsible AI

In 2024, the major code language model, StarCoder2, was trained by Hugging Face and ServiceNow. Crucially, this training utilized a transparent, deduplicated subset of GitHub repositories preserved by Software Heritage.

This successful use case showed that high-quality AI development is achievable while adhering to principles of transparency and responsibility (as outlined in 2023). This trend, now adopted by major players, supports the advancement of open AI models.

Building on that momentum, we launched CodeCommons, an ambitious initiative funded by France 2030. Its goal is to create the world's most comprehensive digital commons for code: a shared, qualified, and traceable foundation enabling responsible, sovereign, and efficient AI. CodeCommons brings together a coalition of research teams dedicated to developing tools for dataset qualification, attribution, transparency, and reproducibility — elements essential for fair and sustainable AI innovation.

The next decade in global infrastructure

Software Heritage has reached a turning point. Our mission, community, and responsibilities — spanning science, culture, cybersecurity, and AI — have all expanded.

To guide our growth, we established the Software Heritage Advisory Board. **This board brings together outstanding experts from Europe, the Americas, and Asia**. Their mandate is to help shape the governance, legal structure, and sustainable financial model needed for Software Heritage to evolve into a fully independent, global, multi-stakeholder organization capable of serving society for generations to come.

This effort is grounded in the trust placed in us by UNESCO, Inria, and our global partners, addressing the urgent need to safeguard the digital knowledge essential to our societies.

Roberto Di Cosmo
Director Software Heritage

A call to action

To everyone who has supported us — researchers, sponsors, engineers, librarians, ambassadors, public institutions, and the global open-source community — thank you. You have helped transform a bold idea into a global public good.

We now call on all stakeholders, academia, industry, governments, international institutions, civil society, from individual contributors to large philanthropy, to join us and continue building this shared infrastructure of knowledge — open, resilient, and universal.

About us

Software Heritage is a non-profit multi-stakeholder initiative launched by Inria in partnership with UNESCO, hosted by the Inria Foundation, and with a growing number of partners. It's building the **universal Archive and knowledge base of software source code**, at the service of society as a whole.

The Software Heritage Archive is the largest collection of publicly available source code ever built. **As of December 2025**, it contains over **27 billion unique source files** from over **421 million projects.**

21,475,236,602	Directories
102,794,272	Authors
133,422,407	Releases
35	Ambassadors
25	Sponsors
20	Team members
1	Advisory Board

HOSTED BY

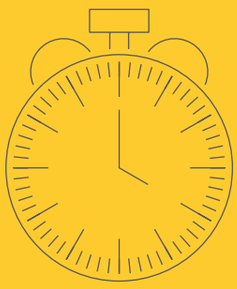
RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Inria

Inria FONDATION

IN COLLABORATION WITH

unesco

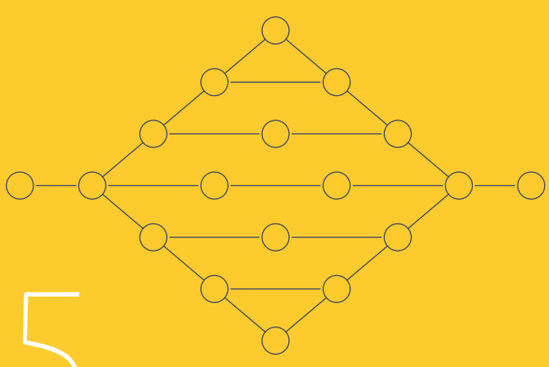


Growing every second since 2015.

85 files archived per second

27

BILLION FILES ARCHIVED



5

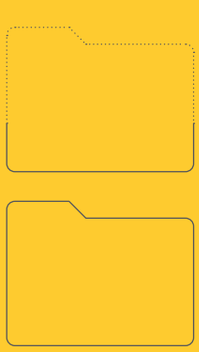
BILLION COMMITS

18 commits per second

1.3 full projects saved per second

421

MILLION PROJECTS SAVED IN TOTAL

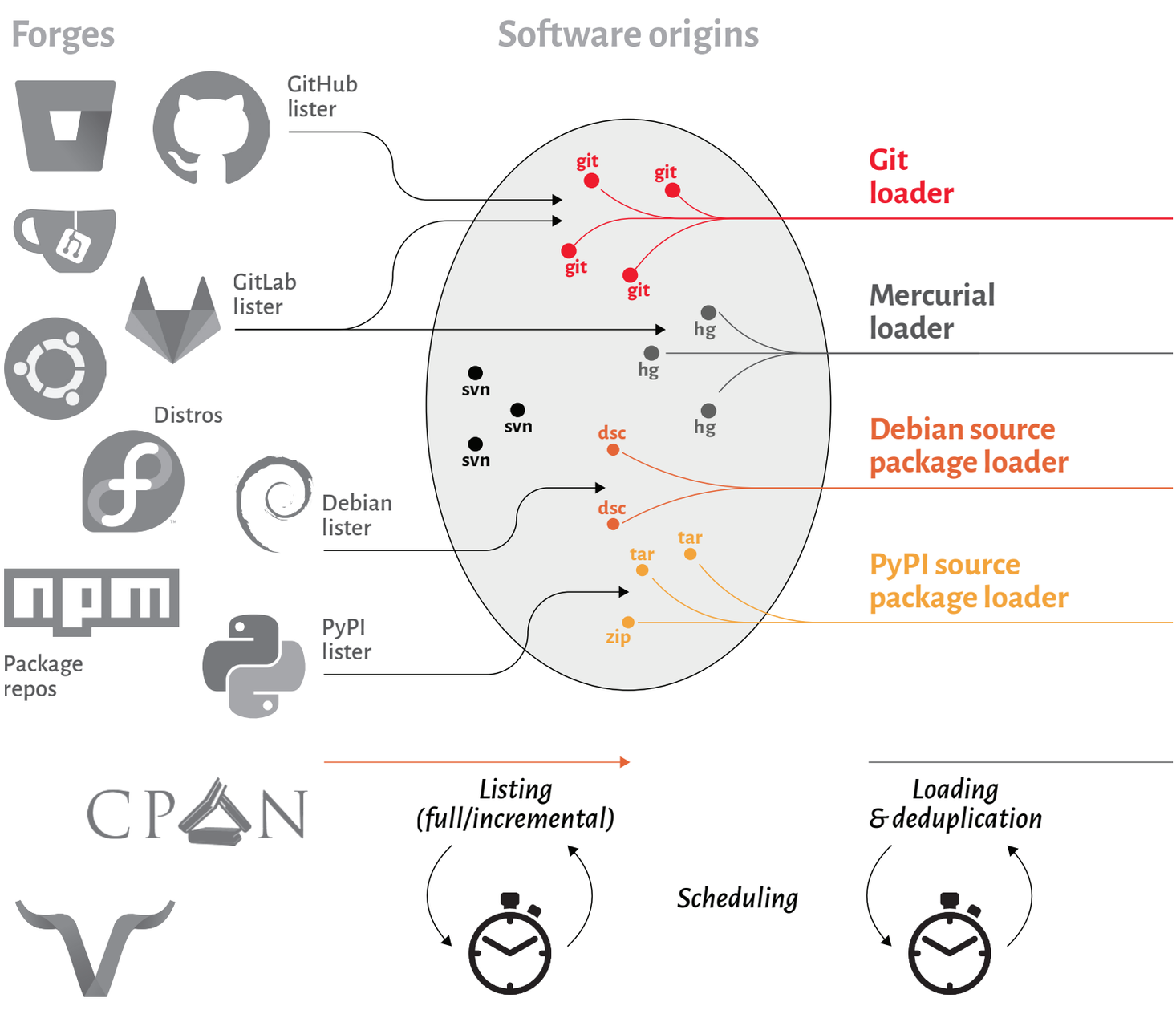


Data: Archive.softwareheritage.org

The largest publicly available social graph

The Software Heritage data structure is a natural extension of Merkle trees, a classical cryptographic construction that combines a code tree with a hash function. (Merkle, 1987)

A massive crawler harvests source code from different sources and converts it, with all its development history, into a single giant Merkle directed acyclic graph, using SWHID cryptographic identifiers for all its nodes.



50 **BILLION NODES**

01 **TRILLION EDGES**

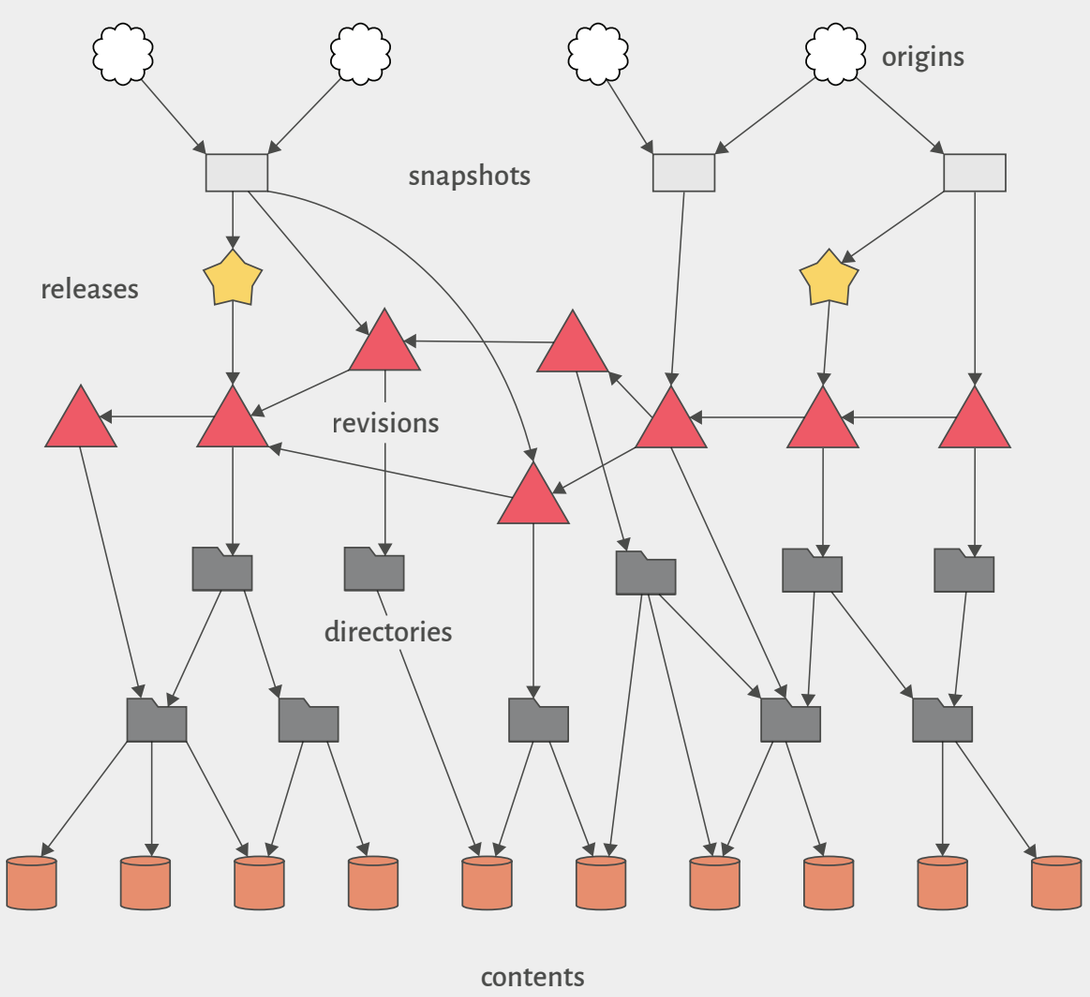
The Software Heritage Archive offers a "single source of truth" that ensures source code remains available forever, stays tamper-proof through cryptographic integrity, and remains fully traceable back to its origin.

Availability: Providing a permanent, resilient archive of the global software commons to prevent data loss and link rot.

Integrity: Guaranteeing authenticity through SWHID cryptographic identifiers to ensure code remains tamper-proof.

Traceability: Utilizing the giant graph to map the provenance, evolution, and reuse of software across all origins.

Software Heritage Archive



Tech highlights

› WHAT'S NEW

Expanding the Archive

To significantly increase Archive coverage, we're preparing to open archiving for several new origins, including crates.io and Maven Central (moving toward production). We've also paved the way to utilize the French high-performance computer AdAstra, increased the current ingestion speed, and enabled the ingestion of repositories with conflicting SHA1 code.

Enriching the Archive

To gather more knowledge about the Archive's content, we began computing new metadata about the source code language, the existence of known vulnerabilities affecting it, and the project's context, such as merge requests, issues, or comments. We also opened a new API to allow partners to send us metadata and connections between code and research articles.

Empowering Archive users

We've added several new features for Archive users.
First, for scientists who want to run their existing analysis tools on parts of the Archive, we improved the efficiency and usability of swh-fuse. This now lets them access the Archive as if it were a local filesystem in HPC environments.

Building on the launch in 2024 of the citation feature, we've also enhanced it to allow scientists to easily reference code in their articles, including the now-standardized SoftWare Hash IDentifier (SWHID) (more on this below) for traceability and identification.

› COMING SOON

Dataset services for large language models

Allow AI trainers to query Software Heritage: "Build a dataset of source code in [Python, Go, C++, ...], with licenses in [MIT, BSD, GPL, ...], updated in the last two years and with no known vulnerabilities," for training or specializing LLMs.

› THE NEXT DECADE

Search engine

A long-awaited feature, improved search will make browsing the Archive easier and help users quickly find the most relevant results.

Infrastructure resilience

Software development relies on infrastructure — forges, package managers, and similar services. Archiving this data is a first step to provide a fallback if an infrastructure becomes unavailable and to prevent data loss. Next steps will increase capacity and add resilience services on top of the archived data.

Gathering comprehensive software knowledge

Software is more than source code and development history; it also includes packages, documentation, and licenses. These can be archived or linked to the current Archive to support compliance, research and preservation.

Software Heritage officially endorsed the UN open-source Principles, solidifying our commitment to a sustainable digital future and aligning our mission with global standards for open, collaborative innovation.

AI coding assistants are quickly becoming indispensable tools for developers. But the provenance of the code they're trained on is often murky, leading to concerns around transparency and author rights. A new initiative launched by the non-profit Software Heritage hopes to change this by providing the world's largest repository of ethically sourced code for training AI.

> ***Edd Gent** - IEEE Spectrum*

CodeCommons

Launched in 2025, the two-year **CodeCommons** project is supported by the French government and French and Italian academic partners. Its goal is to build smaller, higher-quality datasets for **responsible AI tools, leveraging the Software Heritage Archive.** As part of this effort, **CodeCommons** is stress-testing the limits of the swh-fuse system; preliminary experiments on the 10,000-core Kraken cluster at Université Grenoble-Alpes validated a high performance, demonstrating an optimal file storage rate of **30,000 reads per second.**

Mirrors

Software Heritage is strengthening the long-term preservation of code by building a distributed network of mirrors. **Each mirror is a complete copy of the Software Heritage Archive,** operated by a partner institution. This crucial effort prevents information loss and simplifies global access. Building on **ENEA (Italy)**'s existing contribution, we added **GRNet (Greece)** as a new mirror in 2025. **UNIDue (Germany)** is currently finalizing deployment and is expected to join the live network soon, alongside several more organizations.

› FEATURES

Browse & search

The Archive is the gateway to all captured source code and its entire development history. With the browsable platform, it's possible to visualize all visits made to a given location of the code (collected from different forges, package managers and distros) and read the source code content captured.

Deposit

S.W.O.R.D (Simple Web-Service Offering Repository Deposit) is an interoperability standard for digital file deposits. It allows a client (a repository, e.g. HAL) to submit software source archives and metadata to the Archive. Metadata can be also submitted referencing a repository URL (origin) or a SWHID.

Save Code Now

It will take some time to get to every repository in the world, especially if these repositories keep on changing several times a day. This is why the "Save Code Now" service is provided, to give the possibility to notify SWH with a save request.

SWHID provider & resolver

We provide a persistent identifier (PID) that can identify each and every source code artifact with integrity, called a SWHID. SWHIDs are intrinsic identifiers which are intimately bound to the designated object, they do not need a register, only an agreement on a standard to resolve them. The SWHID can also be used as a badge.

Download

The vault is the service in charge of reconstructing parts of the Archive as self-contained bundles, that can then be imported locally. For instance in a Git repository. With the vault, directories and revisions can be downloaded by users on the web platform or through the API. Go to the download vault API reference *https://docs.softwareheritage.org/devel/swh-vault/api.html*

Add Forge Now

In 2022, we introduced a feature called "Add Forge Now", to allow any user to propose archiving of *a whole forge*. The process follows a validation workflow, including curation and verification that the forge technology is supported by Software Heritage tools.

Bulk on-demand archival

A bulk version of the "Save Code Now" feature, which provides a dedicated pipeline for rapid ingestion of a large list of origins, and provides real-time status information for the ingestion progress.

Citation

A new web UI feature enables the generation of software citations, provided that the root directory of the browsed objects contains a citation (.cff or codemeta.json file). The first supported format for generated citations is the BibTeX format.

swh-scanner

The code scanner is a SWHID-based command-line interface tool that compares a local code-base with SWH Archive to identify which artifacts are already known in the Archive, and retrieves information on possible **provenance** (origin of the first occurrence).

Spotlight: CodeMeta
Driving the global metadata effort

Since 2015, CodeMeta has transformed how software is described and cited. By creating a unified schema, it ensures research software is recognized as a first-class scholarly output.

Global adoption & impact
Major international infrastructures—including **Software Heritage, Zenodo, HAL, and SciCodes**—now utilize CodeMeta as the gold standard for interoperability. Its structured mappings allow seamless metadata exchange across the global ecosystem.

Key milestones
- **Version 3.0 & beyond**: Accelerated evolution through the **FAIRCORE4EOSC** and **FAIR-IMPACT** projects.
- **Enhanced tooling:** Launch of the **CodeMeta Generator,** simplifying high-quality metadata creation for researchers.
- **Policy alignment:** Formalized via the **Research Software Metadata Guidelines (RSMD)** to support reproducibility and long-term preservation.

Community governance
CodeMeta is driven by an international cohort of over **40 organizations**, including software engineers, librarians, and standards bodies.

What's next
Our next phase focuses on deepening institutional repository support and aligning with global Open Science ambitions to ensure software remains **visible, citable, and valued.**

CodeMeta continues to evolve with the release of version 3.0, and more recently with the initiative's participation in FAIRCORE4EOSC and FAIR-IMPACT

Spotlight: SoftWare Hash IDentifier (SWHID)

On April 23, 2025, the SoftWare Hash IDentifier (SWHID) reached a historic milestone with its official publication as ISO/IEC 18670, evolving from a foundational design choice within Software Heritage into a globally recognized standard for identifying software source code.

SWHID embodies a core principle of Software Heritage: reference software in a verifiable and decentralized way using intrinsic identifiers, computed directly from software artifacts themselves. This makes them independent of institutions, hosting services, or time, and allows anyone to verify integrity without relying on a trusted intermediary.

Today, tens of billions of software artifacts in the Software Heritage Archive are identified using SWHIDs, ensuring accessibility, guaranteeing integrity, and enabling traceability.

Developed through an open, community-driven process, SWHID provides a durable foundation for making software a first-class component of our shared digital heritage, supporting reproducible science, cybersecurity, supply-chain resilience, and responsible AI.

Learn more at **www.swhid.org**

SWHID provides a durable foundation for making software a first-class component of our shared digital heritage

Open Science highlights

The software pillar of Open Science

In 2025, Software Heritage consolidated its role as the third pillar of Open Science, alongside open access to publications and open research data. Preserving, referencing and connecting research software is now recognized as essential for reproducibility, transparency, and long-term verification. Our operations this year focused on strengthening this pillar through services, partnerships, and community-building that support researchers, institutions, and scholarly infrastructures.

Understanding & aligning with needs

Software Heritage engages deeply with the academic and Open Science ecosystem to understand how researchers, institutions, and infrastructures work with software. Research software calls for specific solutions in scholarly infrastructures, requiring stable identifiers, durable archival workflows, interoperable metadata, and community-driven standards. To align with real needs, we maintain active connections with communities including EOSC, RDA, Force11, French Committee for Open Science, SciCodes, CodeMeta, and others. Recent contributions include showcasing SWHIDs at conferences and webinars, sharing guidance for software preservation, and participating in cross-infrastructure and journal-level discussions on research software citation, metadata, and long-term sustainability.

In the European Open Science Cloud, Software Heritage contributes to strategic reflections on software as research infrastructure, with continued participation in the EOSC Opportunity Area 7: Research Software experts group.

Open Science blueprint
Gruenpeter, M., Di Cosmo, R., Granger, S., Abramatic, J.-F., Cruse, L., & Martinelli, N. (2025). Software Heritage: Open Science Strategic Blueprint. Zenodo. https://zenodo.org/records/17051965

OSPO-RADAR: Building institutional capacity

2025 marked the launch of the OSPO-RADAR project, a two-year initiative supported by the Sloan Foundation to equip universities and research organizations with tools to manage and understand their software assets. Building on the CodeMeta standard recommended by the European Commission for metadata interoperability, we also established the Open Science helpdesk to support the academic community in preserving, citing and discovering software through Software Heritage, and to provide training materials and resources.

https://www.softwareheritage.org/2025/04/02/ospo-radar-project-launch/

A unified view of the research software landscape

FAIRCORE4EOSC and FAIR-IMPACT provided the context for developing bridges between infrastructures for an interconnected and mutualized effort. Making software a first-class output in the research lifecycle.

One lesson stands above all:
Cross-infrastructure coordination succeeds when policy, standards, and community consensus come first—ensuring that technical solutions serve a shared, durable vision of Open Science.

SUCCESS

Multiple infrastructures now reference Software Heritage as the software pillar of Open Science, alongside publications and research data.

IMPACT

EOSC stakeholders see software preservation as foundational for scientific reproducibility, long-term knowledge retention, and interoperability.

Prototyping future workflows through SoFAIR

The SoFAIR project allowed Software Heritage, CORE, HAL, and research partners to prototype next-generation reproducibility workflows by detecting software mentions in articles and connecting to an archival record using the SWHID.

SUCCESS

Combined HAL–SWHID workflows and curated metadata guidelines showed how software, data, and publications can be linked in a reproducible pipeline.

IMPACT

Institutions now have a practical, interoperable blueprint for linking publications to software, improving reproducibility and long-term traceability of research outputs.

Scholarly ecosystem partnerships Deposit & Integration Partners

Deposit Partners: These infrastructures integrate deposit workflows (SWORD v2, Save Code Now, SWHID display, and Browse iframe).
Repositories: HAL (CCSD), InvenioRDM / Zenodo (CERN), DANS / DataverseNL, NII / JAIROCloud (Japan) – deposit service
Publishers: Episciences, eLife, IPOL, Dagstuhl
Aggregators: swMath, OpenAIRE, FR Research Software Catalog – HAL integration, CORE, Research Software Directory.
Reproducibility Platforms: Graphics Replicability Stamp Initiative (GRSI), GNU Guix.

Libraries & Open Science: A strategic alliance

Libraries are pivotal actors in Open Science, and Software Heritage actively collaborates with them to ensure that software receives the same level of preservation, reference, and stewardship as publications and data.
The Archives and Libraries Interest Group (ALIG) brings together libraries and national consortia and individual organizations committed to supporting unfettered access, reference, and citation of software produced by academic research.

Software Heritage meets needs not met by other infrastructures

> ***Frédéric Saby,*** *Deputy Director General, Documentation at Université Grenoble Alpes*

Canadian Research Knowledge Network (CRKN)

01 Memorial University of Newfoundland
02 Université de Montréal
03 University of Toronto
04 York University
05 University of Saskatchewan
06 University of New Brunswick
07 Western University

Consortium of Swiss Academic Libraries (CSAL)

08 Universität Bern
09 ETH Zürich
10 Universität Luzern
11 Université de Neuchâtel
12 CERN

Council of Australian University Librarians (CAUL)

13 University of New South Wales (UNSW)

Couperin

14 Université d'Evry-Val-d'Essonne
15 Université de Haute-Alsace
16 Université de Lille
17 Université de Rennes 2
18 Nantes Université
19 Université de Bretagne Occidentale
20 Le Mans Université
21 Université de Rennes
22 Institut Pasteur

23 DANS
24 Dutch Research Council (NWO)
25 Sant'Anna School of Advanced Studies
26 Université Paris-Saclay

Librarians play a key role in institutional Open Science strategies.

Supporting the supporters

Librarians play a key role in institutional Open Science strategies. Software Heritage supports them through accessible documentation, SWHID guides, CodeMeta resources, and training tailored to software stewardship.

Our participation in events such as the 2025 ADBU Congress reinforced collaborations with French academic libraries and highlighted the growing role of software in library-driven Open Science initiatives.

Learn more in our interview series:

Research & security highlights

From raw code to software knowledge

Di Cosmo, R., Zacchiroli, S. (2023)
The SoftwareHeritage Open Science Ecosystem. In: Mens, T., De Roover, C., Cleve, A. (eds) Software Ecosystems. Springer, Cham.

The Software Heritage Archive is a gold mine of opportunities for applied research on software. As the largest public collection of source code, Software Heritage does more than just Archive: it connects code artifacts developed across different communities and locations worldwide, acting as the essential network map of global software development.

Over the past 10 years, teams from around the world delved into the Software Heritage Archive to conduct cutting-edge research **in fields as varied as software engineering,** cybersecurity, economics, sociology, chemistry, biology, and more.

Read on for a snapshot of research enabled by Software Heritage over the past decade. A full list of research papers co-authored by Software Heritage team members is available on the publications page.

The 2 petabyte problem

Paolo Boldi, Antoine Pietri, Sebastiano Vigna, Stefano Zacchiroli
Ultra-Large-Scale Repository Analysis via Graph Compression. SANER 2020: 184-194.

Antonio Boffa, Roberto Di Cosmo, Paolo Ferragina, Andrea Guerra, Giovanni Manzini, Giorgio Vinciguerra, Stefano Zacchiroli
On the compressibility of large-scale source code datasets. J. Syst. Softw. 227: 112429 (2025).

Tommaso Fontana, Sebastiano Vigna, Stefano Zacchiroli
WebGraph: The Next Generation (Is in Rust). WWW (Companion Volume) 2024: 686-689.

The sheer scale of the Software Heritage Archive — **2 petabytes of source code and over 1 trillion graph edges** — makes comprehensive analysis difficult. Simple data-crunching methods fail outright, and deep analysis remains prohibitively expensive, accessible only to researchers possessing vast computing infrastructures.

Working with the Universities of Pisa and Milan, the Software Heritage research team has developed innovative techniques that enable efficient processing of the Archive's massive content using hardware resources readily available in any standard research laboratory.

Advanced compression resolves the scale challenge for code analysis. The entire graph of public code development now fits directly in the memory of a single large server, making this tool accessible to most labs. Additionally, source code written in the five major languages (C, C++, Java, JavaScript, Python) has been reduced from 78 terabytes to just 3 terabytes. This dramatic reduction in scale democratizes research on the great library of source code for researchers worldwide.

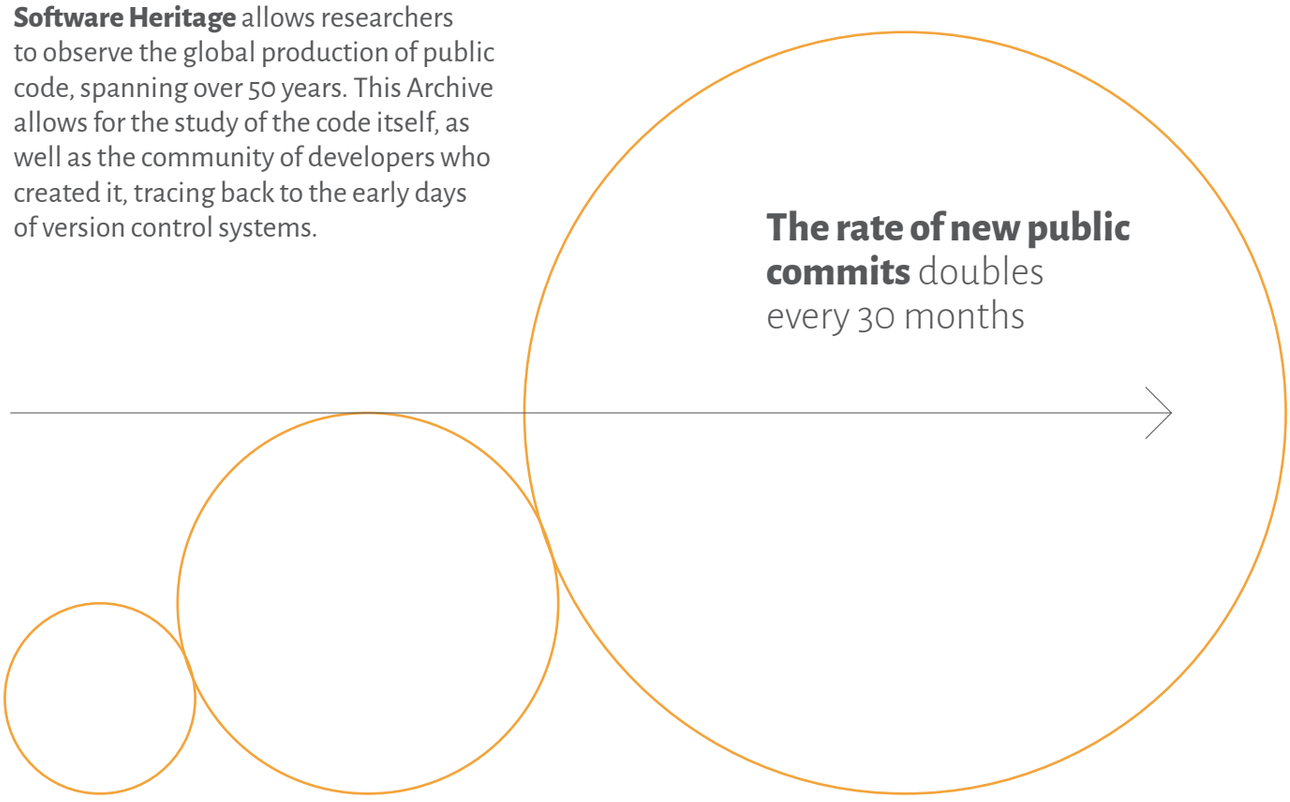
Datasets for the masses

https://datasets.softwareheritage.org

Openness is a defining principle for Software Heritage. True to this commitment, the organization curates and maintains a large body of open datasets that researchers and analysts can freely use to gather insights on the largest public collection of software ever assembled. Users can easily find, retrieve, and correctly credit these resources via our dedicated dataset portal.

Global open-source software production

Software Heritage allows researchers to observe the global production of public code, spanning over 50 years. This Archive allows for the study of the code itself, as well as the community of developers who created it, tracing back to the early days of version control systems.



- **Code acceleration -** Researchers have shown that the global production of public code is growing exponentially: **the rate of new public commits doubles every 30 months** (new public source code files double every 22 months), **a trend that has held steady for over 20 years.** The production of public code has never been faster, and Software Heritage is busy preserving this accelerating history for generations to come.

Guillaume Rousseau, Roberto Di Cosmo, Stefano Zacchiroli
Software provenance tracking at the scale of public source code. Empir. Softw. Eng. 25(4): 2930-2959 (2020)

- **Gender diversity -** Most public code is produced by men, with women accounting for only 8% of the commits archived by Software Heritage up to 2019. The yearly contribution by women had increased steadily, but the COVID pandemic reversed this positive trend. Significant effort is still needed to achieve a more gender-diverse ecosystem of software production, a trend that Software Heritage helps monitor.

Stefano Zacchiroli
Gender Differences in Public Code Contributions: A 50-Year Perspective. IEEE Softw. 38(2): 45-50 (2021).

Annalí Casanueva Artís, Davide Rossi, Stefano Zacchiroli, Théo Zimmermann
The impact of the COVID-19 pandemic on women's contribution to public code. Empir. Softw. Eng. 30(1): 25 (2025).

- **Global contributions -** Code has poured in from all over the world over the past decades. While North America dominated early software development, its dominance later alternated with Europe. In more recent years, the center of gravity has shifted, witnessing the rising importance of contributions from Central and South America and Asia.

Davide Rossi, Stefano Zacchiroli
Geographic Diversity in Public Code Contributions: An Exploratory Large-Scale Study Over 50 Years. MSR 2022: 80-85.

Making open-source software safer for everyone

Open-source software powers the products of daily life, from smart refrigerators to cars. But outdated code is a massive vulnerability. Regulators globally — from the EU's Cyber Resilience Act to US Executive Orders — demand better. Software Heritage is the universal, open knowledge base making open source secure for everyone, guaranteeing availability, integrity, and traceability for the code that powers our lives. The Software Heritage for Security (SWHSec) project brings together eight research teams to leverage information collected by Software Heritage to make open-source software more secure for everyone.

Software Heritage is the map to the stars for software scientists—a complete atlas previous generations could only dream of. For the first time, we can study the software commons as a whole, not merely fragments, enabling reproducible research at unprecedented scale and building software science on solid empirical ground.

> ***Stefano Zacchiroli***
- Chief Scientific Officer, Software Heritage

Know your software... through our lenses

Daniele Serafini, Stefano Zacchiroli
Efficient Prior Publication Identification for Open Source Code.
OpenSym 2022: 12:1-12:8

https://docs.softwareheritage.org/devel/swh-scanner

Modern applications are built by reusing thousands of third-party open-source components, which significantly boosts development productivity. However, this reuse makes Knowing Your Software (KYSW) tricky, requiring due diligence to respect licenses and stay alerted to known vulnerabilities.

Our solution is swh-scanner, which allows software integrators to efficiently scan large code bases and compare them directly against the content of the Software Heritage Archive. The scan results quickly surface the presence of any unexpected third-party components in IT products, tracing their lineage back to where Software Heritage first detected them across open-source ecosystems.

Detecting repository tampering

8.7M

destructive alterations in

1.2 M

public Git repositories

Solal Rapaport, Laurent Pautet, Samuel Tardieu, Stefano Zacchiroli
Altered Histories in Version Control System Repositories: Evidence from the Trenches.
ASE 2025.

Git repositories are central to the software supply chain, serving as the source for open-source code often retrieved through automated workflows. Crucially, these public repositories are not tamper-proof; maintainers or attackers can subtly alter them.

Software Heritage, by providing a comprehensive, periodic Archive, offers the raw data needed to detect these content alterations between archival runs. Researchers from the SWHSec project at Télécom Paris conducted the first large-scale analysis, finding over 8.7 million destructive **alterations** across more than **1.2 million public Git repositories.**

While some alterations are benign (e.g., bot-driven fixes), others are suspicious or nefarious, involving the retroactive removal of sensitive "secrets" or changes to software licenses. The good news: original, pre-alteration content, vital for forensic analysis, can still be retrieved from Software Heritage.

Sector engagement

Industry

Software Heritage is leading the effort to secure the open-source supply chain and address new regulations like the Cyber Resilience Act (CRA). As a founding member of the Open Regulatory Compliance Working Group (ORCWG), it ensures that compliance fosters, rather than hinders, innovation. The infrastructure establishes a clear source of trust through a global Merkle graph, using the SoftWare Hash IDentifier (SWHID) for cryptographically strong tracking of every artifact. This unique capability is leveraged to support expert research teams building cutting-edge cybersecurity tools and provides the only global infrastructure capable of tracking inadvertently leaked code across all distribution platforms.

Academia

Researchers & research software engineers

Research software often disappears or becomes uncitable, undermining verification and impact. Software Heritage ensures long-term preservation, SWHID-based citation, and easy archiving through Save Code Now and CodeMeta tools, reducing code loss and supporting reproducibility for researchers and RSEs.

Libraries & Archives

To help libraries preserve and describe software, Software Heritage provides the essentials: a stable, non-commercial Archive, plus SWHID/metadata training and the ALIG community. This empowers institutions to finally curate and safeguard software in their scholarly collections.

Institutions & OSPOs

We help institutions manage their code outputs. Software Heritage provides the tools—unified visibility, simple deposit workflows, and forthcoming dashboards—to ensure universities meet all Open Science requirements for governance, reporting, and long-term reproducibility.

Scholarly infrastructures

We ensure repositories and journals can guarantee persistent software access. By using our deposit service and SWHID references, publications get durable, software-aware links. This cuts broken links and makes the software behind research reliably citable and verifiable.

We should consider the efforts of Software Heritage. Located in France, they're archiving the entirety of software created globally, and we need a European initiative to build upon this.

> ***President Emmanuel Macron***
> *- "Choose Europe for Science" launch*

Culture & education

We produced five new, product-centric videos in partnership with UNESCO. They deliver concise content covering our key offerings: software deposit in HAL, SWHID, save code services, and the Software Heritage Acquisition Process (SWHAP).

Watch the series online:

Public administration

Software Heritage serves as the central, long-term Archive for public software, enhancing transparency and improving citizen services through assured preservation and accessibility. France's Interministerial Directorate for Digital Services (Dinum) exemplifies this by systematically archiving all open-source software developed by its national public administrations. By depositing software and machine-readable metadata in the Archive, public bodies guarantee the availability and reuse of their digital assets for future generations.

Collaboration & community

Presence in National and International Policies

Software Heritage is now referenced in major policy instruments shaping Open Science:

▸ **French Second National Plan for Open Science (PNSO2)**, which identifies software as a core pillar and highlights the need for long-term preservation.

▸ **French National Roadmap for Research Infrastructures**, where Software Heritage is positioned as a critical component for software-related scholarly infrastructure.

▸ **European Open Science Cloud (EOSC)** strategic documents, especially through Opportunity Area 7 on Research Software, where SWH contributes to defining priorities for metadata, PIDs, and archival workflows.

Contributions to policy & standards

Software Heritage actively informs policy development at both national and international levels:

▸ Participation in **Open Research Community Working Groups (ORCWG) and Collective Rights Action (CRA)** initiatives, helping define next-generation norms for research software management, reuse, and recognition.

▸ Technical and conceptual contributions to the interpretation and deployment of the **EU Copyright Directive**, ensuring that preservation, access, and reuse of research software remain legally supported in cultural, scientific, and educational contexts.

Projects

Software Heritage collaborates in a wide portfolio of European, international, and philanthropic projects that advance FAIR and open research infrastructures:

EU-funded projects: FAIRCORE4EOSC, FAIR-IMPACT, SoFAIR (CHIST-ERA), and contributions across the EOSC framework.

Philanthropic and foundation-funded initiatives: Sloan Foundation OSPO-RADAR project supporting research software management in universities.

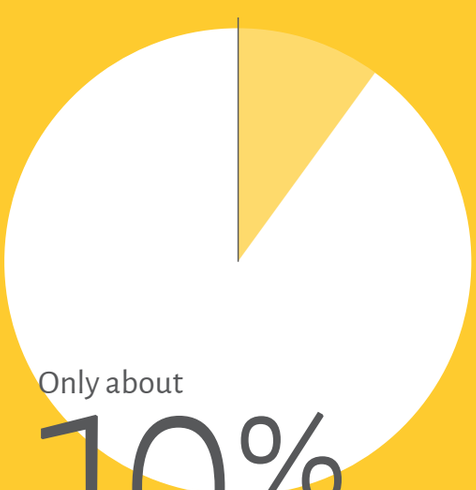
NGI (Next Generation Internet) collaborations: advancing trustworthy, open, and decentralized digital infrastructure aligned with long-term software preservation needs.

Spotlight: CERN

The full scope of open-source contributions

Software Heritage and CERN teamed up to track down one fundamental thing: the true scale and location of all CERN-related code.

The challenge is clear: only about 10% of the "official" open-source code comes from employees. The other 90% is created by visiting researchers whose work is scattered across various platforms and repositories before they move on. Getting a complete, accurate picture of this community's massive impact is a challenge, but it's essential work. Look for full analysis and key findings from the year-long research project in 2026.



Only about

10%

of "official" open-source code from employees

Before this analysis, we had no idea of the true scale of CERN's decentralized contributions. The outcome showed a massive scale—over six million commits—and a remarkable breadth. We found not just scientific software, but popular general-purpose packages that built thriving communities well beyond CERN. The real value of Software Heritage is its unified scope across all platforms. This allows us to comprehensively measure CERN's impact, proving that our contribution to the open-source ecosystem extends far beyond traditional academic citations

> ***Micha Moskovic***
> *INSPIRE Product Manager, CERN*

Meet the Advisory Board

Started in 2025, **this new board will guide Software Heritage through a substantial transition**. Their crucial work includes four core areas: identifying the right legal framework, developing a sustainable financial model, building our autonomous team, and strengthening global partnerships.

Jean-François Abramatic
› INRIA scientist emeritus

Cat Allman
› VP, Open Source, Digital Science

François Bancilhon
› Entrepreneur, Inria

Claudia Bauzer Medeiros
› Professor of Computer Science, UNICAMP

Jean-Pierre Bourguignon
› Member of the Committee on Publishing of the International Mathematical Union

Paul Ginsparg
› Professor of Physics and Information Science, Cornell University

Wayne Graham
› Chief Information Officer, CLIR

Mike Milinkovich
› Executive Director, Eclipse Foundation AISBL

Simon Phipps
› Standards & Policy Director, OSI

Laurent Romary
› Director of Scientific Information and Culture, Inria

Aurélie Simard
› Executive Director, Inria Center of Expertise for International Cooperation on AI

Oscar Valenzuela
› Principal Open Source Engineer, Amazon

Peter Wang
› Chief AI and Innovation Officer and Co-founder, Anaconda

Kazu Yamaji
› Professor and Director of Research Center for Open Science and Data Platform, National Institute of Infomatics (NII), Japan

Meet our team

EXECUTIVES

Roberto Di Cosmo
› Founder, CEO

Stefano Zacchiroli
› Founder, CSO

Thomas Aynaud
› CTO

VISITING SCIENTIST

Mathilde Fichen

OPEN SCIENCE

Sabrina Granger
› Academia Engagement Progam Manager

Linda Angulo Lopez
› Open Science partnerships specialist

ENGINEERS

Renaud Boyer
Nicolas Dandrimont
Thibault Deregnaucourt
Antoine R. Dumont
Martin Kirchgessner
Ulysse Kuchler
Antoine Lambert
Valentin Lorentz
Amadou Thiam
Aymeric Varasse

MANAGEMENT

Benoit Chauvet
› Project Manager

David Douard
› Dev Team Manager

Morane Gruenpeter
› Director of Scholarly Ecosystems

Bastien Guerry
› Head of Partnerships, Public Sector & Industry

Nicole Martinelli
› Strategic Communications

Vincent Sellier
› Sysadmin Team Manager

ADMIN AND EVENTS

Aurélie Morin
› Office Manager

Marla da Silva
› Events

© Inria / Photo B. Fourrier

Ambassadors

At Software Heritage, we understand that success is achievable only through the collective efforts of a diverse community. Since 2020, our ambassador program has played a crucial role in nurturing collaboration and promoting adoption. This year, we added four new advocates with a wide range of geographical and technical expertise.

Agustín Benito Bethencourt
› Independent Consultant

Alex Khrustalev
› Principal Software Engineer

Alexis Lebis
› Assistant professor in computing, IMT Nord Europe

Anna-Lena Lamprecht
› Professor, Software Engineering, Postdam University

Baptiste Mélès
› Research Associate, CNRS

Bariş Günör
› Junior software developer

Bertrand Néron
› Research Engineer Bioinformatics, Institut Pasteur

Bostjan Spetic
› Head of collections & research, Slovenian Computer Museum

Bruno Khélifi
› Staff researcher, IN2P3

Camille Françoise
› Program Manager & Policy Advisor, Member of the board of trustees of Wikimedia France

Cécile Arènes
› Head of Research Data & Digital Humanities, Sorbonne University Library

Flavia Marzano
› President of the Scientific Committee, Ampioraggio Foundation

Florent Zara
› Open Source Services Team Lead, Eclipse Foundation

Frédéric Santos
› Data analyst, CNRS

Gavin Henry
› Ant Networks Founder

Giacomo Lorenzetti
› Research Software Engineer, CEFCA

Harish Pillay
› Leader of the Community Architecture and Leadership in Red Hat Asia Pacific

Italo Vignoli
› Founding member: Document Foundation & LibreOffice project

Jaime Arias
› Research Engineer, CNRS

Joenio Marques da Costa
› Research Software Engineer, Gustave Eiffel University

Julien Caugant
› Assistant librarian, Aix Marseille University Library

Malin Sandström
› Senior Research Officer, Swedish Research Council

Max Kalik
› Software Engineer, Triumph Labs

Maxence Azzouz-Thuderoz
› Data scientist in AI and natural language processing, FIZ Karlsruhe - Leibniz Institute

Mohammad Akhlaghi
› Staff researcher, CEFCA

Neha Oudin
› Data Platform Engineer, Canonical

Nika Maltar
› Software Archive and Collection Coordinator Technisches Museum Wien

Océane Valencia
› Head of archives and records department, Sorbonne University Library

Pierre Poulain
› Associate professor in Bioinformatics Université Paris Cité

Sandrine Layrisse
› System and network administrator, CNRS

Shiraz Malla Mohamad
› Junior data analyst

Simon Delamare
› Research engineer, CNRS

Simon Phipps
› Computer engineer, independent consultant

Violaine Louvet
› Research Engineer, Grenoble Alpes University

Wendy Hegenmaier
› Head of Software Preservation and Emulation, Yale University

Become an Ambassador

Interested in becoming a **Software Heritage Ambassador ?**
Tell us about yourself and your interest in our mission:

ambassadorprogram@softwareheritage.org

Gaining alignment through community engagement

We build alignment by **connecting communities that use, publish, preserve, and curate software**. Through workshops, working groups, and collaborations, we create shared understanding and foster adoption across disciplines and infrastructures.

A key example is the annual **Software Heritage Community Workshop**, which brings together institutions, researchers, librarians, publishers, and infrastructures to discuss use cases, workflows, and future needs.

For Software Heritage's 10th anniversary, a landmark exhibition at UNESCO will celebrate source code as a cultural, scientific, and artistic artifact.

Easing adoption

We disseminate knowledge through a growing **learning ecosystem**, including tutorials, training materials, guides, and ambassador-led sessions. These resources support reproducible research practices, metadata curation, and software preservation workflows:

We also support "cultural brokers"—ambassadors, developers, maintainers, librarians, researchers, and OSPOs—who bridge communities and help integrate Software Heritage practices locally.

The UNESCO Code Exhibition

For Software Heritage's 10th anniversary, a landmark exhibition at UNESCO will celebrate source code as a cultural, scientific, and artistic artifact. Highlighting historic programs, research software, and community-shared stories, the exhibit reveals how code shapes our world across science, industry, and society. Visitors will explore unique features from the Software Heritage Archive — Save Code Now, reproducibility tools, acquisition workflows, and software stories — while discovering contributions from librarians, developers, scientists, artists, and students. Fully open, participatory, and international, the exhibition invites everyone to share how their code reflects heritage, creativity, and human knowledge.

Sponsors

› DIAMOND SPONSORS

IBM

Microsoft

TII Technology Innovation Institute

› PLATINIUM SPONSORS

HUAWEI

cnrs

Infosys

cea

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

› GOLD SPONSORS

aws

Google

Université Paris Cité

openinventionnetwork

› SILVER SPONSORS

AdaCore

Canadian Research Knowledge Network
Réseau canadien de documentation pour la recherche

GitHub

Ministère des Armées
Liberté
Égalité
Fraternité

DGA
Direction générale de l'armement

Premier ministre
Liberté
Égalité
Fraternité

Direction interministérielle du numérique

Università di Pisa

Sorbonne Université

› BRONZE SPONSORS

Sant'Anna
Scuola Universitaria Superiore Pisa

NWO

université PARIS-SACLAY

Konsortium der Schweizer Hochschulbibliotheken
Consortium des bibliothèques universitaires suisses
Consorzio delle biblioteche universitarie svizzere
Consortium of Swiss Academic Libraries

Scuola Normale Superiore

Université de Lorraine

Become
a sponsor:

Software Heritage

A decade of milestones