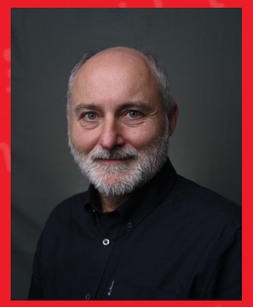
Software Heritage

Membership Program Source Code & AI

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Collecting, preserving and sharing software source code since 2015

“Our sponsors enable Software Heritage’s mission: to collect, preserve, and share source code. They stand at the crossroads of culture, research, and technology, safeguarding it as a precious artifact of our digital age, while empowering discovery and innovation for generations to come.”



Roberto Di Cosmo
Co-founder & CEO
Software Heritage

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 is building the **universal archive and knowledge base of software source code**, at the service of society as a whole.

The Software Heritage Symposium and summit 2024 took place at UNESCO’s headquarters on February 1st bringing together advisors, sponsors, ambassadors, and the entire community.
© Inria | Photo M. Magnin.

The **Software Heritage archive** is the largest collection of publicly available source code ever built, containing, as of November 2024, more than 21 billion unique source files from over 335 million software origins.

Hosted by

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Inria

Inria La Fondation

In collaboration with

unesco

Software Heritage has been launched by Inria in 2015.

Our mission

SOFTWARE HERITAGE IN A NUTSHE

We are building an essential infrastructure, that is meant to ensure three main properties for the source code we collect:

Availability

The code will be stored, preserved and made accessible on the long term.

Traceability

Eeach software component will get a unique identifier, called **SWHID**, that can be relied upon in the long term.

Uniformity

Despite the great variety of origins, all of the source code collected in our archive will be accessed through the same uniform Application Programmer's Interface (**API**)

A catalog to find them all

Software is spread all around: it is developed on many collaborative platforms and distributed through a variety of different channels. Software Heritage is building a **universal catalog** to let you **find** all software projects, no matter where they are developed, or how they are distributed.

An archive to preserve them

Modern software development relies on collaborative platforms, and many of them can be used free of charge. One can **create**, but also **modify** or **delete** projects: *they are not archives*. In recent years, we have seen several platforms come and go, sometimes suddenly, endangering hundreds of thousands of software projects all at once. Software Heritage is building the **universal archive** that is needed to ensure we will not loose source code any more.

© iStock | kynny

An instrument to explore and study them

Software underlies all aspects of our modern societies, and in a few decades we have built software systems of incredible complexity: some are huge programs, with tens of millions of lines of code, some are smaller programs, but mostl rely on hundreds or thousands of other components. We need to master this complexity, in order to build better, safer systems, and protect against malware.

Humankind has been able to build marvelous instruments to explore the universe, now it's time to build a common, shared infrastructure to explore and study the galaxy of software development. With enough support, Software Heritage can evolve into such an infrastructure.

HUMANKIND HAS BEEN ABLE TO BUILD MARVELOUS INSTRUMENTS TO EXPLORE THE UNIVERSE, NOW IT'S TIME TO BUILD A COMMON, SHARED INFRASTRUCTURE TO EXPLORE AND STUDY THE GALAXY OF SOFTWARE DEVELOPMENT.

TECHNOLOGY HIGHLIGHTS

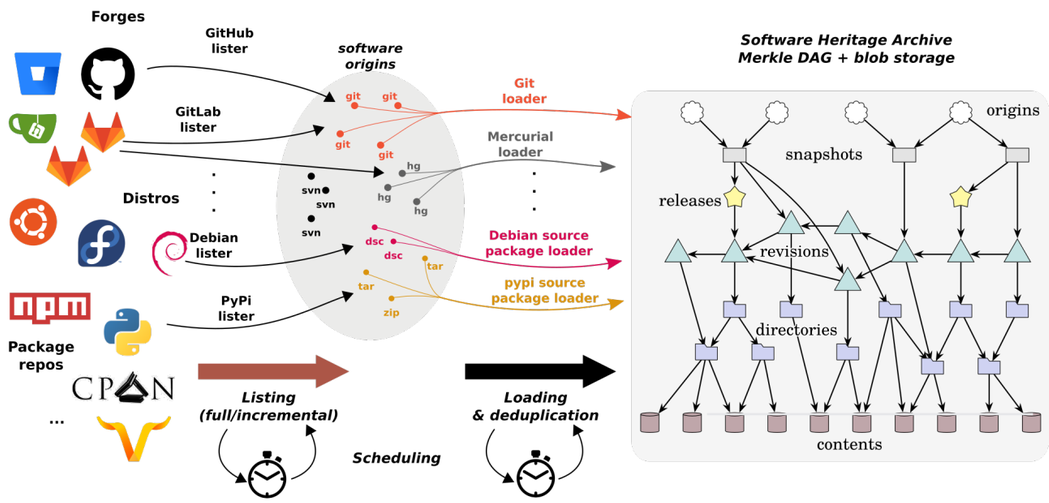
All the public code history in a giant graph

Merkle graphs and SWHID

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 | **700** billion edges

THE SOFTWARE HERITAGE DATA STRUCTURE IS A NATURAL EXTENSION OF MERKLE TREES, A CLASSICAL CRYPTOGRAPHIC CONSTRUCTION, COMBINING A TREE AND A HASH FUNCTION. [MERKLE, 1987]



The process is separated into three phases: ***listing software sources***, ***scheduling updates*** and ***collecting the software artifacts*** into the archive.

Towards transparency in AI

By promoting transparency and responsible stewardship, Software Heritage aims to help researchers, developers, and organizations navigate the challenges of AI in code-based applications.

Software Heritage provides the Software Hash Identifier (SWHID) for over 50 billion software artifacts it has collected from over 300 million projects, ensuring availability, guaranteeing integrity and enabling traceability of all its contents.

Publishing a list of SWHID identifiers of source code used in training datasets contributes to transparency in AI.

The SWHID intrinsic persistent identifier

All artefacts in the Software Heritage archive get a **S**oft**w**are **H**ash **ID**entifier, or **SWHID** for short, that is guaranteed to remain stable (persistent) over time.

A SWHID consists of two parts, a mandatory *core identifier*, and an optional list of *qualifiers* that specify the context and can pinpoint a subpart.

Software Heritage Statement on Large Language Models for Code

As we strive to preserve this vital resource for future generations, we acknowledge the emergence of inquiries regarding the use of the Software Heritage archive for the training of machine learning models, particularly large language models (LLMs) that can automatically generate code to assist with software development tasks.

In alignment with our mission, we believe that LLMs for code should be built in a transparent and respectful way, to the benefit of all. We hence state the following principles for acceptable machine learning use of the Software Heritage archive.

Principles

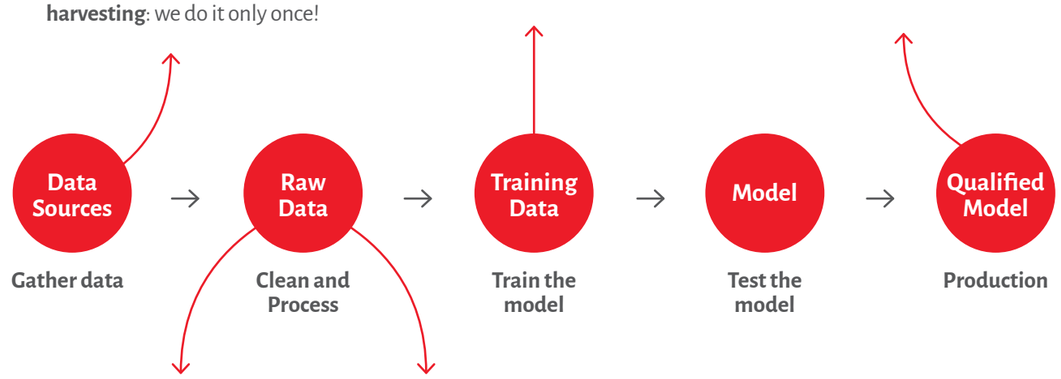
1. Knowledge derived from the Software Heritage archive must be given back to humanity, rather than monopolized for private gain. The resulting machine learning models must be made available under a suitable open license, together with the documentation and toolings needed to use them.
2. The initial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

A Step Forward: Code Commons

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

Traceability: Identify and make available the data used for training

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



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

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

Supporting Software Heritage

Becoming a member

One common infrastructure shared

Software Heritage builds a key digital shared infrastructure:

- It needs your support, because it is a non-profit operation;
- As a digital commons, you mutualize the cost with other sponsors.

Your contribution supports

- **Infrastructure Development:** storage, compute, and network resources to support the ever-growing archive.
- **Software and Standards Development:** new key features to enhance archive functionality and usability.
- **Innovation and Alignment:** stay at the forefront of innovation and aligned with the latest advancements (like AI and Cybersecurity).
- **Community Engagement:** foster adoption, grow the community and connecting with partners worldwide.
- **Open Source Preservation:** a safe haven for Open Source software, safeguarding it against loss and making it perpetually retrievable.
- **Operational Sustainability:** infrastructure expansion, software development, and normal operations to keep the archive and the organization running smoothly.
- **Long-Term Mission Support:** bring the infrastructure to operational and financial maturity, ensuring the long-term sustainability of operations.

Benefits

1. Supporting Software Heritage's long term mission contributes to your **social responsibility agenda**
2. You have an opportunity to **contribute your views on the technical roadmap**
3. You get **advanced information on new developments**
4. The annual general meeting is a great occasion to **meet the supporter network** and learn about the latest progress and perspectives
5. Depending on local legislation, your contribution **may be tax deductible**

Source Code and AI Interest Group

The **Source Code and Artificial Intelligence Interest Group** brings together stakeholders committed to advancing the mission of the Software Heritage initiative by developing Artificial Intelligence tools that enhance access to the archive and simplify the use of its content. These efforts, such as the creation of open Large Language Models, aim to benefit the common good.

Membership tiers

Members can choose to join at different tiers, depending on the level of support they want to provide for the stability, long term availability and evolution of the Software Heritage archive infrastructure.

Diamond members

See Software Heritage as a strategic infrastructure and commit significant resources that contribute to the ***stability of the platform*** in the medium to long term, as well as to its necessary ***technological development.*** They contribute to the ***strategic roadmap*** of Software Heritage.

Platinum members

Consider Software Heritage as a core infrastructure and provide funding for its ***regular operation*** as well as advice for its short term development. They contribute at different levels, according to tier, to the ***technical roadmap*** of Software Heritage.

Gold members

See Software Heritage as an important infrastructure for enabling transparent and responsible AI, and want to ***get involved in the community*** that sustains it.

Fees and Benefits

Software Heritage is a Foundation, hosted by the Inria Foundation.
Memberships and associated benefits valid in the period 2024-2025 are summarised below:

Commitment of 1 year or 3 years (recommended)	Diamond	Platinum			Gold
		T3	T2	T1	
Annual Membership and Benefits*	**> 250 K€**	**> 200 K€**	**> 150 K€**	**> 100 K€**	**> 50 K€**
Strategic Advisory Board	✓				
Contribute to Platform Stability	✓	✓			
Technical Advisory Board	✓	✓	✓		
AI Working Group	✓	✓	✓	✓	
General Assembly	✓	✓	✓	✓	✓
Code Deposit for Models Transparency	✓	✓	✓	✓	✓

(*) a reduced academic fee may be available for qualifying institutions, inquiry to learn more.

Join us!

Becoming a sponsor is very easy: contact us at sponsor@softwareheritage.org or fill the online form

You are in great company!
Thrive together!

Diamond Sponsors

Software is key in **CEA**'s commitment to transferring knowledge from research to industry.
With the Software Heritage Foundation, we stand behind the preservation and sharing of this knowledge.

Platinum Sponsors

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

The National Open Science Plan was launched on 4 July 2018 by the Minister of Higher Education, Research and Innovation. This plan includes a provision to support Software Heritage, an initiative that we consider a major pillar of open science. In addition to enabling open access to publications and research data, making research software source code openly available is critical to success of the open science program that we are collectively building.

CNRS's support to Software Heritage, a universal, open and sustainable software archive, is a natural part of our proactive approach in favour of open science, a necessary revolution in which everyone must play a part.

Microsoft has been involved in open source initiatives by enabling, integrating, releasing and contributing to many open source projects and communities for well over a decade. We applaud the Software Heritage as an open project that will help curate and conserve human knowledge in the form of code for future generations as well as help today's generations of developers find and re-use code worldwide.

Intel has been at the forefront of open source development for nearly two decades and today is a top contributor to the Linux kernel, as well as dozens of leading projects across technology markets and industries.
Intel is committed to support Software Heritage in its mission to collect, preserve and share code, as we believe open source is critical in transforming our world through innovation in enterprise, consumer technology, the Internet of Things and beyond.

Huawei has been working with the open source communities for decades: we are active contributors in projects ranging from the Linux kernel to cloud native computing and machine learning, and we will keep increasing our participation and investment in this open innovation world. We share Software Heritage's vision that publicly available source code, including open source software, is a precious heritage of mankind, and should be collected, preserved and shared for the benefit of all.

Gold Sponsors

Google is proud to support Software Heritage in its mission to collect, preserve, and share software for future generations. We look forward to the variety of services that can be built atop this unique collection of software.

Partnering with Software Heritage was a great journey for BigCode and **Hugging Face**. The foundation's focus on preservation, reproducibility, availability and traceability mirrors many of the values and mission of Hugging Face as a central platform for sharing and collaborating in the ML community.

Open source software has been one of the instrumental, driving forces of innovation this century. Software Heritage is an important organization for software, (...). Archiving of code in a curated form maintains the technical and scientific knowledge that goes along with the code, preserving the innovation while also providing a means for determining prior art.

At **ServiceNow** we recognize the value and importance of preserving open-source software (...). We firmly believe in the capacity of Software Heritage to cultivate goodwill and collaboration within the technology ecosystem, while promoting a more sustainable and open software industry.

We are aware of the code's value for our digital transformation, it has become a major asset for the bank and we firmly believe that we must preserve it in the long term. Open Source lies at the heart of our strategy, as it is in line with our needs and our values: team spirit, innovation, responsibility and commitment to better serve our clients.

Firmly committed to open science, which is at the heart of its project, **Sorbonne University** supports Software Heritage. By helping to collect and to share software, Software Heritage contributes to one of the key missions of the university: the preservation and transmission of knowledge and of our scientific heritage.

By supporting the Software Heritage initiative, Université de Paris continues its commitment to the free and responsible sharing of knowledge and research software.

Silver Sponsors

La DINUM

Bronze Sponsors

Ready to make a difference?
Join us!
Together, let's ensure the legacy of technology thrives for generations to come.

Software Heritage will provide solid, common foundations to serve the different needs of heritage preservation, science, and industry.

softwareheritage.org
@swheritage
@SwHeritage
@swheritage@mstdn.social
@softwareheritage4978

Becoming a sponsor is very easy: **fill the online form**

Contact us at: sponsor@softwareheritage.org