Annual Report 2022

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collecting, preserving and sharing software source code since 2015

Software [is our] Heritage

We collect and preserve software in source code form, because software embodies our technical and scientific knowledge and humanity cannot afford the risk of losing it.

Software is a precious part of our cultural heritage. We curate and make accessible all the software we collect, because only by sharing it we can guarantee its preservation in the very long term.

Foreword

On June 30, 2016, after almost two years of preparatory work, we opened up to the world the Software Heritage website unveiling our long-term mission to collect, preserve and make easily available the source code of all software publicly available. We could immediately see how timely was the creation of Software Heritage: we were just starting up, and already busy saving hundreds of thousands of software projects endangered by the demise of Google Code and Gitorious.org.

On April 3rd, 2017, a landmark partnership agreement was signed between UNESCO and Inria to establish a framework for collaboration on preserving the knowledge embedded in software source code and making it widely available, centred around the ongoing development of the Software Heritage archive.

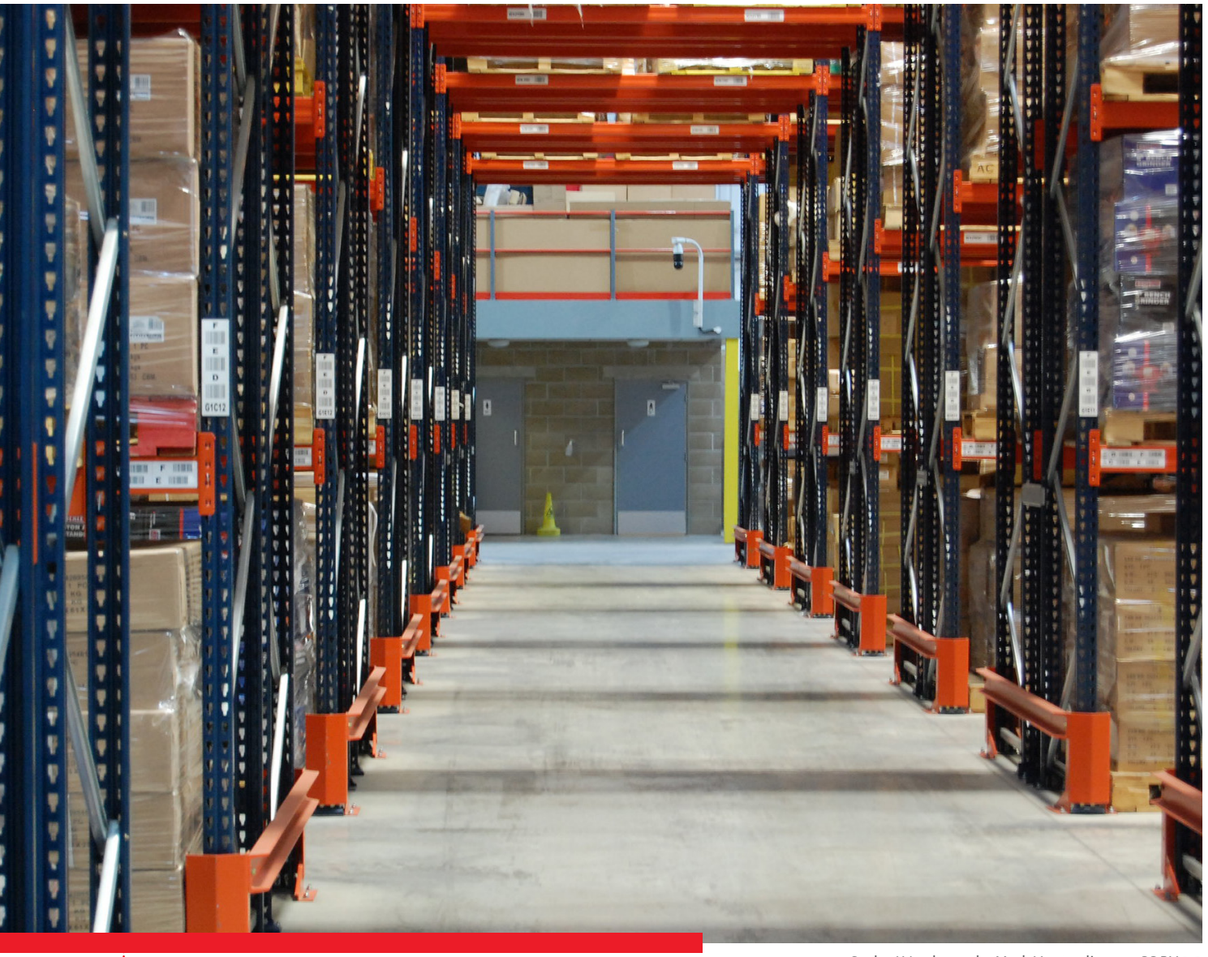
In 2021, we celebrated five years of continuous and passionate dedication to this mission, thanks to the many partners that have made this possible: sponsors that have followed up on Inria's initial full support, UNESCO for the precious collaboration on the vision and the policy issues, private foundations that have made it possible to fund key expert contributors, many organizations and individuals that share the vision, and a growing number of contributors and donors.

This report provides an overview of what Software Heritage does, how it does it, and why it is so important to build an open, non-profit, multi-stakeholder, long-term and common infrastructure to collect, preserve and share all of humankind's source code.

It is a humbling undertaking, and we will only succeed with the contribution of a large community, from funders to users and adopters, from industries to public bodies and individual contributors.

We call today on all stakeholders to take part in Software Heritage's long-term mission.

Roberto Di Cosmo
Co-founder & CEO
Software Heritage



Sealey Warehouse by Mark Hunter, license: CC BY 2.0
https://www.flickr.com/photos/toolstop/4324416999/

WE HARVEST PUBLICLY AVAILABLE SOURCE CODE FROM MANY SOFTWARE PROJECTS AND KEEP UP WITH DEVELOPMENT HAPPENING THERE. AS OF TODAY OUR ARCHIVE ALREADY CONTAINS AND KEEPS SAFE FOR YOU:

- **12,676,800,814** Source files
- **2,679,877,947** Commits
- **182,088,638** Projects

TABLE OF CONTENTS

Learn about the last five years of Software Heritage in just five minutes!

https://youtu.be/Ez4xKTKJO2o

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.

Cultural Heritage

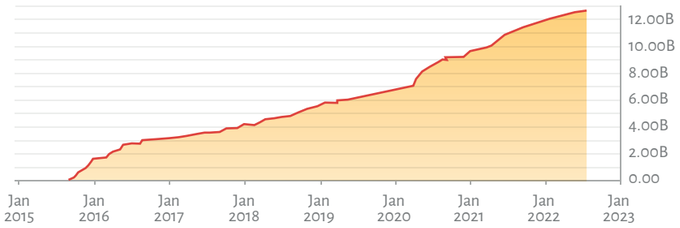
Industry

Reserach

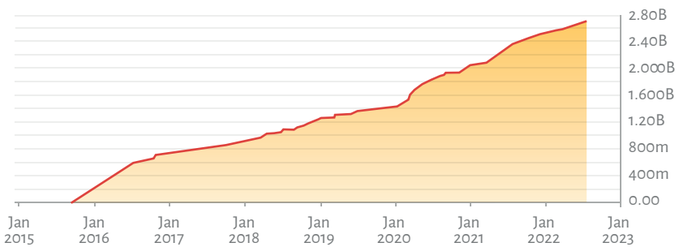
Public Administration

Software Heritage

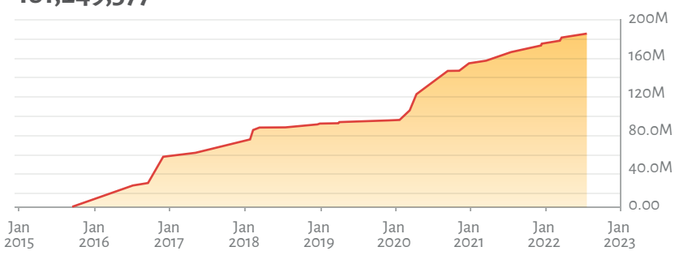
Source Files
12,538,666,608



Commits
2,645,066,174



Projects
181,249,577



Directories
10,342,140,231

Authors
48,778,458

Releases
33,580,610

4	14	20	10	20
Advisors	Team members	Ambassadors	Grantees	Sponsors wordlwide

*On November 30th, **a special event took place at UNESCO's headquarters** to celebrate the five years of Software Heritage.*

The **Software Heritage archive** is the largest collection of publicly available source code ever built, containing, as of July 2022, over 13 billion unique source files from over 194 million software origins.

Hosted by

Inria — Inria La Fondation

In collaboration with

unesco

Software Heritage has been launched by Inria in 2015.

Sponsors

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

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.

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.

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

Open source software has been one of the instrumental, driving forces of innovation this century. Software Heritage is an important organization for software, having already archived more than 6 billion unique source files. 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.

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

A longstanding, stable software repository like the Software Heritage is of direct interest to us, since our aerospace and defense customers are often responsible for projects lasting several decades.

Open Data is the fuel, and Software is the engine of the digital transformation that is driving change in all aspects of modern societies.

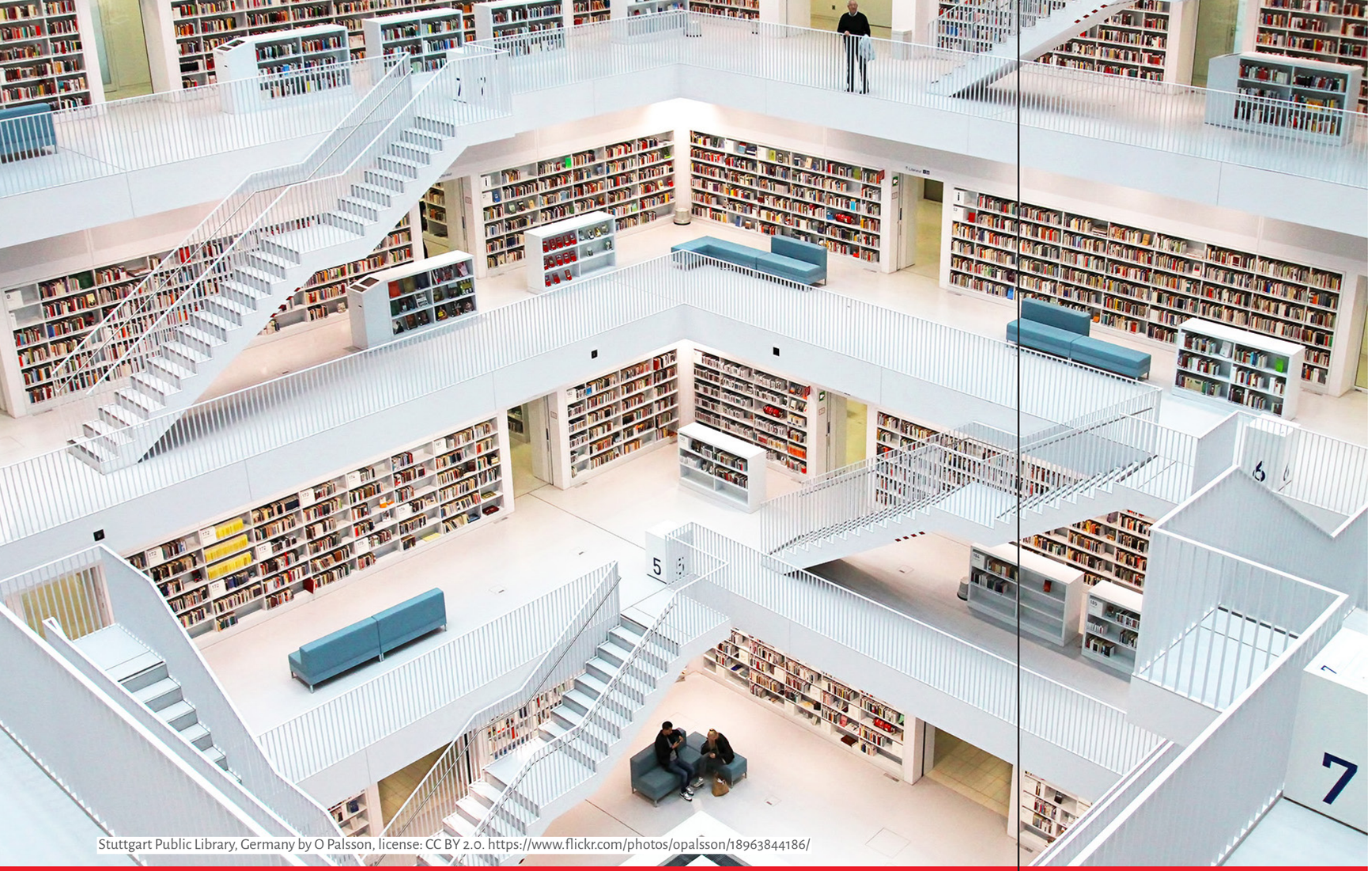
GitHub recognizes the crucial importance of open source software and the work of millions of developers, around the globe, collaborating together. We at GitHub see it as part of our responsibility to protect OSS now and for decades to come.

Free and open source software has always been a vital part of **Google**, as we use and contribute to thousands of projects. It encourages the development of innovative new technology, and opens doors for budding engineers all over the world.

Pisa has been the first Italian University to foresee the future relevance of computer science, and it has committed to its development early on: by supporting the building of the first computer designed in our country in the late Fifties of the last century, and by introducing the first course in Informatics in Italy in 1969.

To collect, preserve, and share all software that is publicly available in source code form is a tremendous task and a tremendously important one.
VMware believes that open source software is an essential building block of today's technology solutions and IT strategies.

Bronze Sponsors



Stuttgart Public Library, Germany by O Palsson, license: CC BY 2.0. https://www.flickr.com/photos/opalsson/18963844186/

OUR MISSION

Our ambition is to collect, preserve, and share all software that is publicly available in source code form. On this foundation, a wealth of applications can be built, ranging from cultural heritage and education to industry, from science to public administration, and more.

"Programs are written for people to read, and only accessorily for machines to execute"

— Harold Abelson

Collect

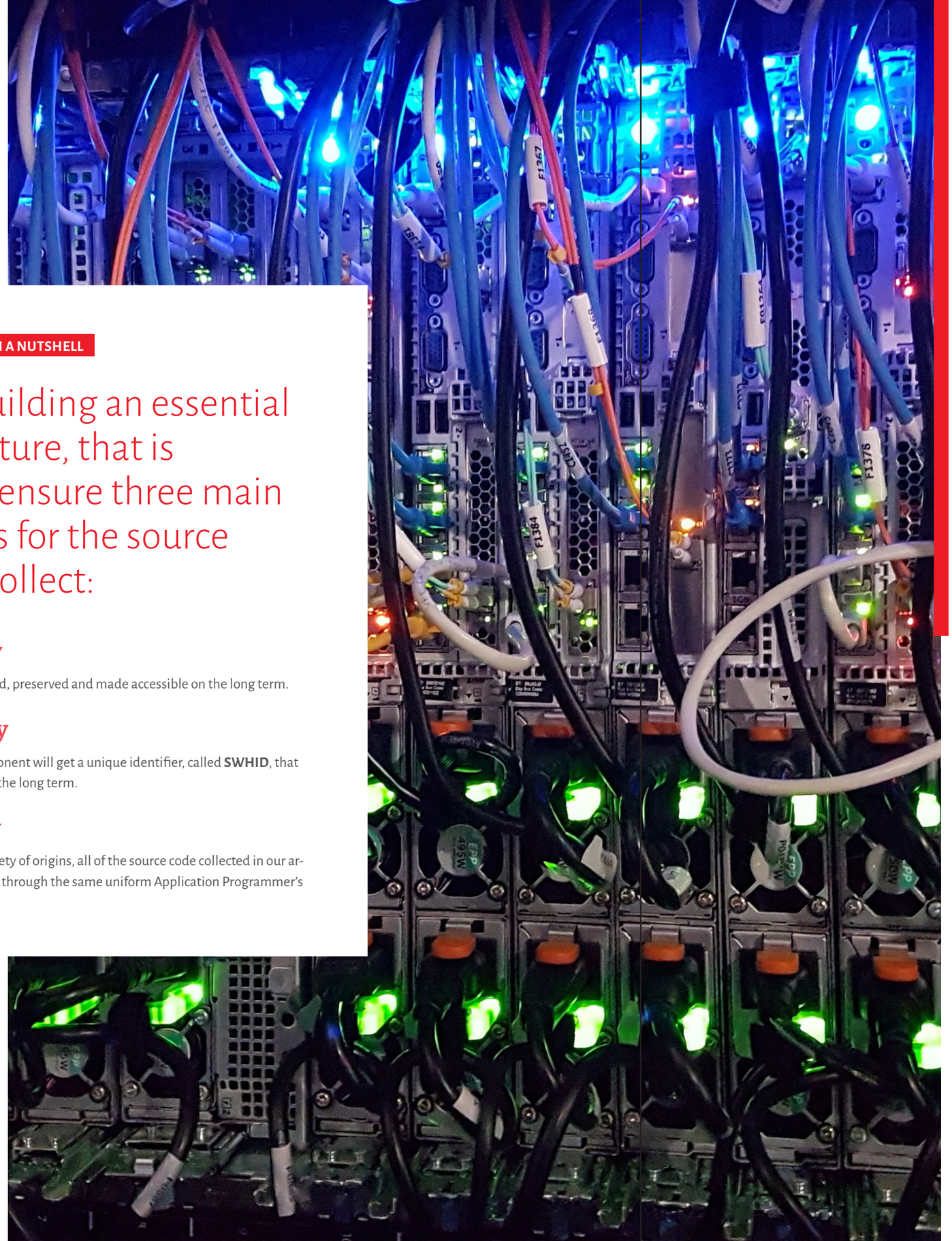
Software is the fabric that binds together our digital lives. Any software component may turn out to be essential in the future, so we **collect all software** that is publicly available in source code form, and we will encourage the construction of **curated archives** on top of Software Heritage.
We keep track of the **origin of software** we archive and store its full development history: this precious meta-information will be carefully harvested and structured for future use.

Preserve

Software is fragile and we are unfortunately starting to lose it, sometimes massively, when popular code hosting platforms shut down or reduce operations. We preserve software, because **it contains** our technical and scientific knowledge. We preserve software because **it is the means of accessing** all of our knowledge. We know that for this to be sustainable, a **vast collective effort** is needed, and we will release as **free/open source software** all the software we write for the needs of Software Heritage and openly describe our technical architecture and processes.
We **are building** an open **network of peers** and mirrors that share with us the responsibility of maintaining several copies of all the software we collect.

Share

We are building the largest archive of software source code ever assembled. We will **index**, **organize**, **make referenceable and accessible** all of this precious heritage.
We provide **the SWHID unique identifiers**, intrinsically bound to the software components, and that need **no central registry**, to ensure that a resilient web of knowledge can be built on top of the Software Heritage archive.
A variety of services, ranging from documentation to classification, from search to distribution, will progressively be developed to release all the potential of this **Library of Alexandria of Software**.



SOFTWARE HERITAGE IN A NUTSHELL

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 **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.

An instrument to explore and study them

Software underlies all aspects of our modern societies, and we have built in a few decades software systems of incredible complexity: some are huge programs, with tens of millions of lines of code, some are smaller programs, but 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.

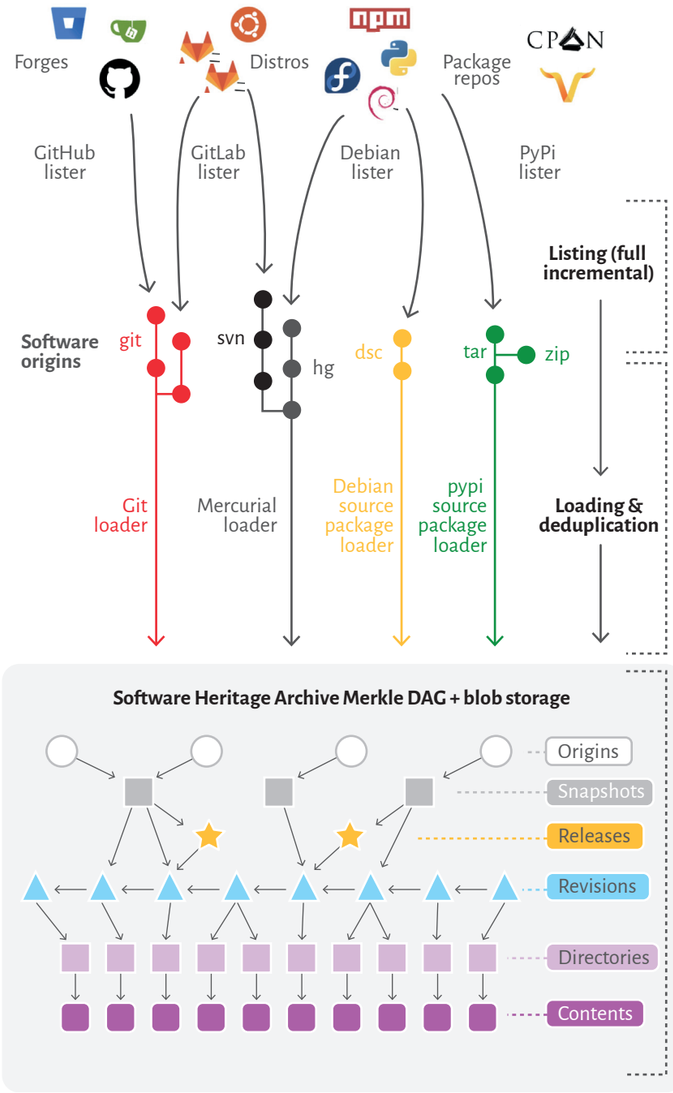
HUMANKIND HAS BEEN ABLE TO BUILD MARVELOUS INSTRUMENTS TO EXPLORE THE UNIVERSE, NOW IT'S TIME TO BUILD A COMMON, SHARED INFRASTRUCTURE TO EXPLORE AND STUDY THE GALAXY OF SOFTWARE DEVELOPMENT.

TECHNOLOGY HIGHLIGHTS

Merkle graphs and SWHID
A giant Merkle graph

A massive crawler **harvests** source code from different sources and converts it, **with all its development history**, into a giant Merkle-directed acyclic graph that contains over **30 billion nodes and 350 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].



Software Heritage Listers

A lister is a software component used for discovering all software projects available on a code hosting or distribution platform. In 2022 we have unveiled a dedicate page with all the available listers and links to their high-level documentation: https://docs.softwareheritage.org/user/listers/

Software Heritage Loaders

A loader is a software component used to ingest a software artifact into the *Software Heritage* archive, performing the appropriate conversion into the Merkle graph.
In 2022 we have unveiled a dedicate page with all available loaders and links to their high-level documentation: https://docs.softwareheritage.org/user/loaders/

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

A COMMUNITY EFFORT

To help cope with the many different technologies out there, we rely on expert contributors, some of which could be funded through cascading grants provided by the **Alfred P. Sloan foundation** and the **NLNet foundation**.

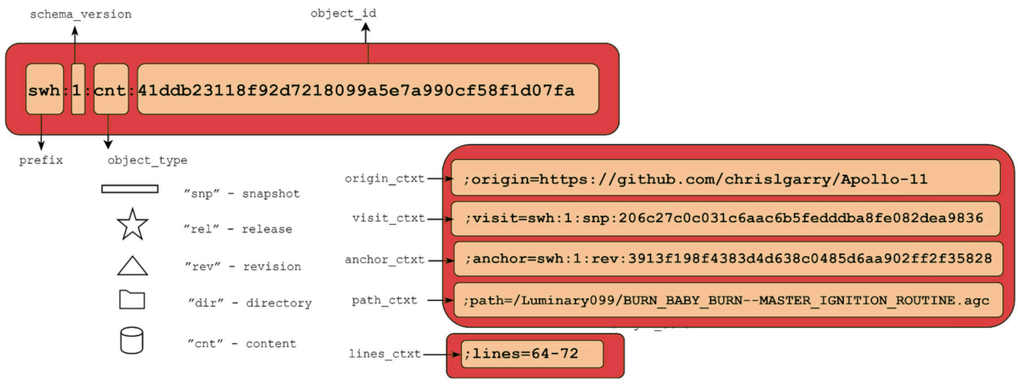
ALFRED P. SLOAN FOUNDATION

nlnet

https://www.softwareheritage.org/grants/

The SWHID intrinsic persistent identifiers

All artefacts in the Software Heritage archive get a **SoftWare Heritage persistent IDentifier**, 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. One can obtain them using the Permalinks sidebar present on all pages of the Software Heritage archive, and the core identifier can be computed independently by everybody.

Learn more about SWHID

https://www.softwareheritage.org/2020/07/09/intrinsic-vs-extrinsic-identifiers/

Intrinsic and Extrinsic identifiers

WHAT MAKES SWHIDS SPECIAL IS THAT THEY DO NOT DEPEND AT ALL ON THE VERSION CONTROL SYSTEM

Building a solid web of knowledge that lasts over time is of paramount importance. A key component of this web are the links between the different entities, that are designated using systems of identifiers that come in two broad categories:

- **Extrinsic**: use a *register* to keep the correspondence between the identifier and the object (e.g. URLs, DOIs)
- **Intrinsic**: intimately bound to the designated object, they do not need a register, only agreement on a standard (e.g. git cryptographic hashes)

The software development world has long ago adopted intrinsic digital identifiers, like git hashes, that enable decentralized operations and independent integrity verification. What makes SWHIDs special is that they do not depend at all on the version control system: any software artifact ingested in the Software Heritage archive gets these identifiers.

SWHIDs are now part of the SPDX 2.2 industry specification, and have corresponding properties in Wikidata. A normalization process is underway.

Services that Software Heritage offers today

Browse & Search

The SWH archive is the gateway to all captured source code and its entire development history. With the browsable platform, it is possible to visualize all the visits made to a given location of the code (collected from different forges, package managers and distros) and read the source code content captured.
https://archive.softwareheritage.org/

SWHID provider & resolver

SWH provides 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. For more information.

Go to the resolver API endpoint https://docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html

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.
For more information
Go to the download directory API endpoint

https://archive.softwareheritage.org/browse/vault/

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.

Go to API endpoint
https://save.softwareheritage.org/

Deposit

The deposit feature is a SWORD 2.0 Server implementation. **S.W.O.R.D** (**S**imple **W**eb-Service **O**ffering **R**epository **D**eposit) is an interoperability standard for digital file deposit. The deposit allows a client (a repository, e.g. HAL) to submit software source archives and its associated metadata to the SWH archive. Metadata can be also submitted referencing a repository url (origin) or a SWHID.

For more information
https://deposit.softwareheritage.org//

NEW IN 2022

Add Forge Now

In 2022 new introduced a feature called "Add Forge Now", to allow any user to propose the archival of *a whole forge*. The process follows a validation workflow, including curation, and verification that the forge technology is supported by Software Heritage tools.
https://docs.softwareheritage.org/devel/swh-lister/tutorial.html#lister-tutorial

Want to find out what's next?
The 2022 technical roadmap is online!
https://docs.softwareheritage.org/devel/roadmap/

Building for the long term

Building a global infrastructure to stand the test of time is a humbling undertaking. To this end, we rely on the following founding principles

Non Profit and Multistakeholder

Experience shows that a single for profit entity, however powerful, does not provide sufficient durability guarantees in the long term. We believe that it is essential to build a **non profit multi-stakeholder** foundation that has the mission of Software Heritage as its explicit primary objective, and we are delighted to be working with UNESCO towards it.

Transparency of code and architecture

Long-term preservation efforts cannot be based on black boxes that hide the process behind closed doors. We are long-time Free/Open Source Software developers and advocates, and our code and specifications are released under a **Free and Open Source Software** license. We are designing a complex software architecture. Its design and specifications are public.

Collaborative development

The mission of Software Heritage is a humbling undertaking: to succeed, a large collective effort is needed. To foster it, we adopt an **open development** process, and strive to create an active community around all components of the Software Heritage infrastructure.

Facts and provenance

Following best archival practices, Software Heritage will store full provenance information, in order to be able to always state **what** was found **where** and **when**.

Intrinsic unique identifiers

Each software component is assigned a **SWHID intrinsic unique identifier** that can be directly computed from and is intrinsically bound to it. It does not rely on third parties, so it is truly persistent, and everybody can build on it.

Mirrors

Any data infrastructure faces multiple challenges over time, that can be technical, organizational or legal.
To minimize the risks over the long term, we are working to build a resilient system. Due to the nature of the archive, we follow a **centralized and replicated** approach, establishing a network of independent full **mirrors** of the archive, but we also look at **decentralized** technologies.

ENEA
Italian National Agency for New Technologies, Energy and Sustainable Economic Development

A shared infrastructure for multiple stakeholders

Culture and education

"[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive"
— Paris Call on Software Source Code[1]

Cultural heritage is the legacy of physical artifacts and intangible attributes of a group or society that are inherited from past generations, maintained in the present and bestowed for the benefit of future generations.
Software in source code form is produced by humans and is understandable by them; it is a special form of **knowledge** that is at the same time **human readable and machine executable**.
It is an important part of our heritage that we cannot afford to lose. Software is furthermore a key enabler for preserving other parts of our cultural heritage that we would de facto lose if we lose the software needed to access them. Preserving software is essential for preserving our cultural heritage.
We have the privilege to be able to talk to most of the people that created this new science and technology of computing, but there we have little time left: it is **urgent** to take action, and Software Heritage is providing guidance and tools, in addition to the archive infrastructure itself.

1 Available from the UNESCO website as ark:/48223/pf0000366715, 2019.

Software Heritage Acquisition Process

Rescuing landmark legacy source code

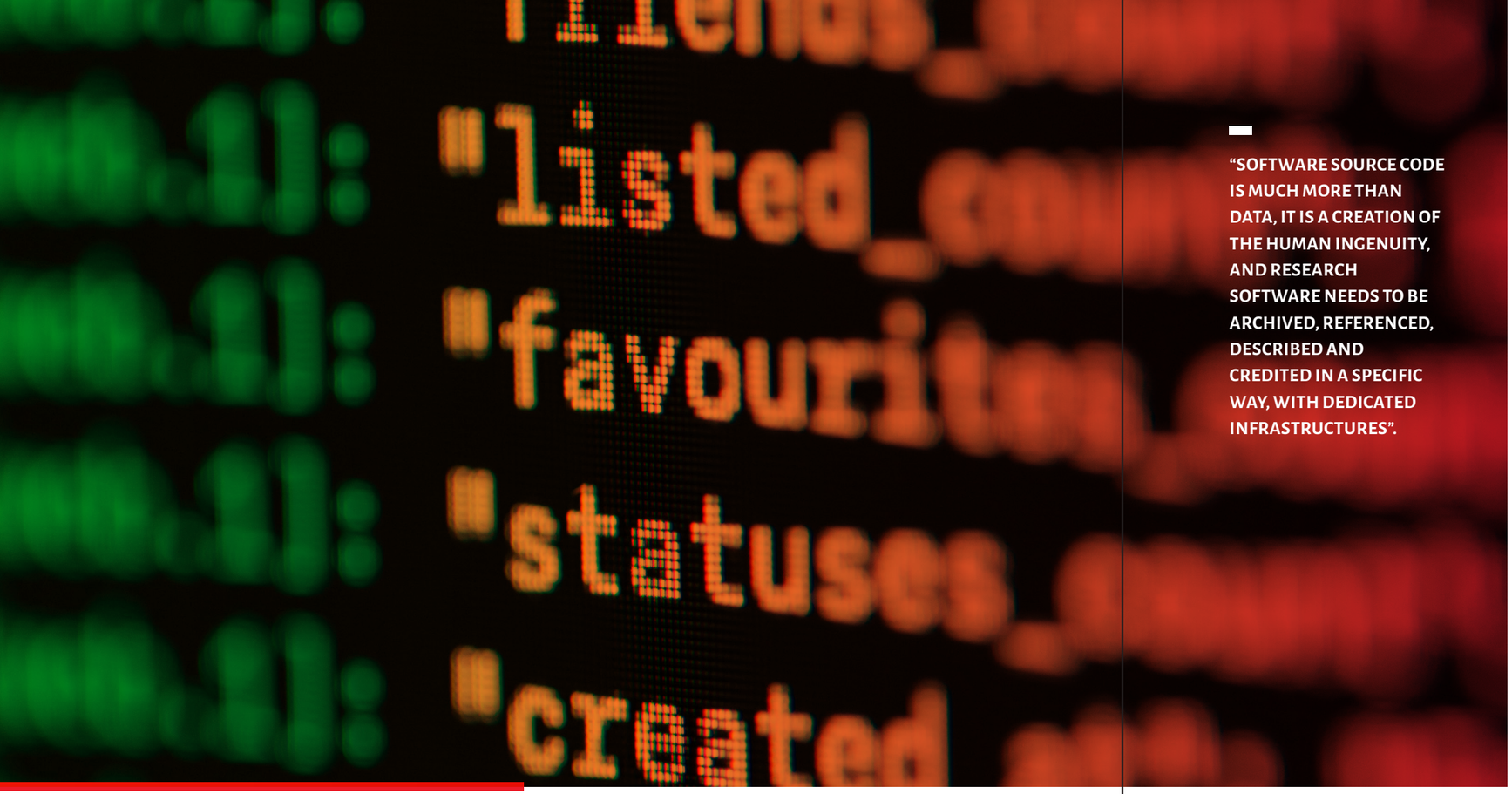
The Software Heritage Acquisition Process (**SWHAP**), developed in collaboration with UNESCO and the University of Pisa, details all the steps needed to successfully curate landmark legacy source code and archive it in Software Heritage.

"THE SOFTWARE STORIES SYSTEM ALLOWS USERS TO CREATE A MULTIMEDIA OVERVIEW OF A LANDMARK LEGACY SOFTWARE TITLE, MAKING IT ACCESSIBLE TO A WIDE RANGE OF SOFTWARE ENTHUSIASTS WITHOUT ANY TECHNICAL BACKGROUND".

Software Stories

Highlighting the human side behind the software projects

Software Stories is a project supported by UNESCO as part of the shared mission to collect, preserve and share source code as a precious asset of humankind.
The Software Stories system allows users to create a multimedia overview of a landmark legacy software title, making it accessible to a wide range of software enthusiasts without any technical background.



"SOFTWARE SOURCE CODE IS MUCH MORE THAN DATA, IT IS A CREATION OF THE HUMAN INGENUITY, AND RESEARCH SOFTWARE NEEDS TO BE ARCHIVED, REFERENCED, DESCRIBED AND CREDITED IN A SPECIFIC WAY, WITH DEDICATED INFRASTRUCTURES".

The Software Pillar of Open Science

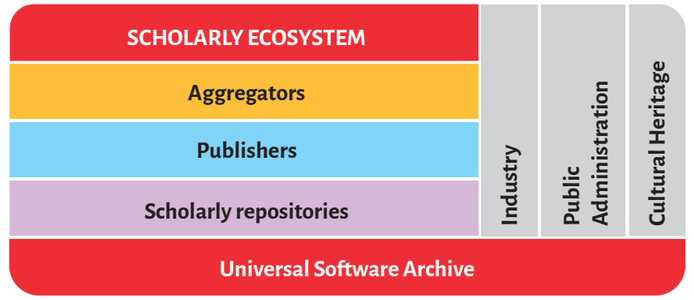
Open science refers to the unhindered dissemination of results, methods and products from scientific research. It draws on the opportunity provided by recent digital progress to develop open access to publications and - as much as possible - data, source code and research methods. Making software source code available, with the option of modifying, reusing and disseminating it, is a major requirement to ensure the reproducibility of scientific findings and to support the creation and sharing of knowledge, in keeping with the open science ethos.

– French second national plan for Open Science, July 2021

Software has become a pillar of research, ubiquitous in all its fields: a large part of the technical and scientific knowledge that is being developed today is described in the **software software code** at a level of detail that is often needed to remove ambiguities that may exist in intuitive descriptions. The preservation of this universal body of knowledge is as essential as preserving research articles and data sets. In the quest to make research results **reproducible**, and pass knowledge to future generations, we must preserve these **three main pillars**: research **articles** that describe the results, the **data** sets used or produced, and the **software source code** that embodies the logic of the data transformation.

Archive, reference, describe and credit research software

Software source code is much more than data, it is a creation of the human ingenuity, and research software needs to be archived, referenced, described and credited in a specific way, with dedicated infrastructures.



As recognized in the EOSC SIRS report published in 2020, Software Heritage provides the core shared infrastructure that allows to interconnect the scholarly ecosystem with all the other ecosystems that rely on software, and provide uniform, long term archival and the SWHID intrinsic persistent identifiers.

A multi-year collaboration between the french national open access portal HAL and Software Heritage has led to developing a seamless workflow to archive, reference, describe and cite research software, and the Second National Plan for Open Science now recommends that all french researchers use it *and fixes the objective to standardize* the SWHID identifiers. *In February 2022, Software Heritage has been inscribed in France in the national roadmap of research infrastructures.*

Running software, again and again

Software Heritage ensures the availability and traceability of software source code, a key prerequisite for reproducing, reusing and adapting existing software. Partnership are being established to connect Software Heritage with package managers and build systems, to enable the replication of full blown executables and systems, the ultimate goal of reproducibility.

Towards a global infrastructure for *research on software source code*

The Software Heritage Graph Dataset

The Software Heritage Graph Dataset is a fully deduplicated Merkle DAG representation of the Software Heritage Archive. The dataset links together file contents identifiers, source code diretories, Version Control System (VCS) commits tracking evolution over time, up to the full states of VCS repositories ass observed by Software Heritage during perioddic crawls. The Dataset's contents coome from major development forges (including GitHub and GitLab), FOSS distributinos (e.g., Debian), and language-specific package managers (e.g., PyPI).

We publish a relational representation of the full archive of Software Heritage as a set of tables. Available as open data in the AWS Open Dataset collection, it makes it easier for researchers to perform large-scale reproducible software studies.

Here is a sample query to find the most popular commit verbs across all the archive.

Query

```
SELECT COUNT(*) AS C, word FROM (
   SELECT word_stem(lower(split_part(
    trim(from_utf8(message)),' ', 1)))
   AS word FROM revision
   WHERE length(message) < 1000000)
WHERE word != ''
GROUP BY word
ORDER BY C
DESC LIMIT 20;
```

Results

Results (20) — Copy — Search rows

#	c	word
1	271573294	updat
2	163328012	merg
3	140044381	add
4	105800317	fix
5	103646653	ad
6	52891401	bump
7	50067041	initi
8	45609622	creat
9	42633225	remov
10	32230842	chang

Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli
The Software Heritage Graph Dataset: Public software development under one roof
In: Proceedings of the 16th International Conference on Mining Software Repositories, pp. 138-142, IEEE Press, 2019.

Traversing the Software Heritage graph at full speed

With ***over 25 billion nodes and over 350 billion edges***, the Software Heritage graph is one of the largest public social graphs available. Thanks to bleeding edge graph compression technology, it can now all fit in 200Gb of memory, and be traversed in ***just tens of nanoseconds per edge***!
Java and gRPC APIs available:
https://docs.softwareheritage.org/devel/swh-graph/grpc-api.html

Paolo Boldi, Antoine Pietri, Sebastiano Vigna, Stefano Zacchiroli
Ultra-Large-Scale Repository Analysis via Graph Compression
SANER 2020, 27th Intl. Conf. on Software Analysis, Evolution and Reengineering. IEEE

The Software Heritage Licence Dataset

6.5 million unique full texts of free and open source license variants extracted from the Software Heritage archive, with example origin repository, and oldest public commit of origin.

Stefano Zacchiroli
A Large-scale Dataset of (Open Source) License Text Variants
MSR 2022 (best dataset paper award)

Key findings

What is the growth rate of public source code?

The amount of original commits in public code doubles every ~30 months and has been doing so for 20+ years; original source code files double every ~22 months.

Rousseau, Di Cosmo, Zacchiroli
Software Provenance Tracking at the Scale of Public Source Code.
In Empirical Software Engineering, 2020.

Diversity, equality, inclusion in public code

Metadata in the archive can be used to study long-term trends of diversity in software development contributions. For example, male authors contributed 92% of public code commits up to 2019.
The ratio of female authors (and their contributions) has grown stably for 15 years reaching for the first time 10% of yearly contributions in 2019, but the COVID-19 pandemic has reversed the trend.

Zacchiroli. **Gender differences in public code contributions: a 50-year perspective.** IEEE Software, 2021

Rossi and Zacchiroli. **Worldwide gender differences in public code contributions.** ICSE SEIS, 2022

Rossi and Zacchiroli. **Geographic diversity in public code contributions.** MSR 2022

Detecting project *forks*

Today, developers contribute to open-source projects by working on their own copies, called ***forks***, that can be created in many ways. The Software Heritage archive allows to detect "exogenous" forks across multiple platforms.

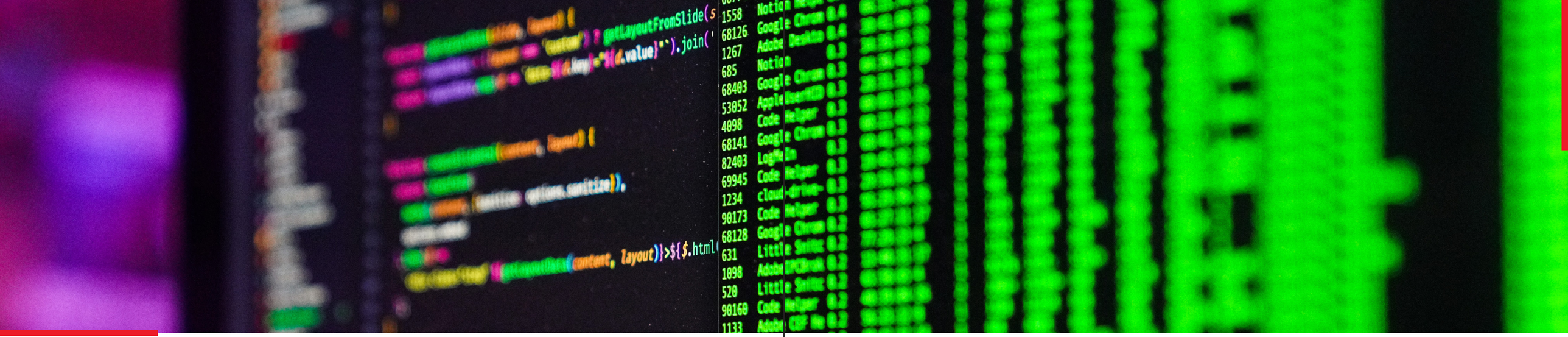
Pietri, Rousseau, Zacchiroli.
Forking Without Clicking: on How to Identify Software Repository Forks. MSR 2020

Awards and recognition

Antoine Pietri, best French PhD in Software Engineering **"Enabling Big Code analysis on a very large source code corpus"**. Awarded by the CNRS research working group GPL.
https://theses.hal.science/tel-03515795v1

Stefano Zacchiroli with **Davide Rossi**. Google Award for Inclusion Research 2022, for the research project **"What Causes the Lack of Diversity in Open Source?"**.
https://research.google/outreach/air-program/recipients/

Antoine Pietri with **Stefano Zacchiroli**. Award Best Dataset Paper: **"A Large-scale Dataset of (Open Source) License Text Variants"**.
https://arxiv.org/abs/2204.00256



Industry

[There is a need to] *ensuring and attesting, to the extent practicable, to the* ***integrity and provenance of open source software used within any portion of a product.***

Executive Order on Improving the Nation's Cybersecurity
White House, May 12, 2021

The ability to use, understand and evolve the processes and devices that run our industry relies on the ability to access, understand, and evolve the software that controls them.
Software Heritage provides a *neutral*, *common*, *shared*, *open*, *non-profit*, **reference knowledge base** encompassing all the software source that is publicly available, enabling new applications to improve all aspects of the software process.

Securing the Open Source software supply chain

The *uniform*, *technology-neutral global Merkle graph* provides, with the growing mirror network, **a transparent source of trust.** The SWHID provides *uniform*, *technology-independent*, *and cryptographically strong* **intrinsic identifiers** to *track source code artifacts at all levels*.
These unique features of Software Heritage ensure **availability**, guarantee **integrity** and enable **traceability** of the source code of all artefacts in the open source software supply chain.

Complete and corresponding source code

Industry players can **delegate** to the Software Heritage archive the task to preserve and make available to third parties the complete and corresponding source code of any open source component they use. They can use the Software Heritage as a trusted reference in their agreements.

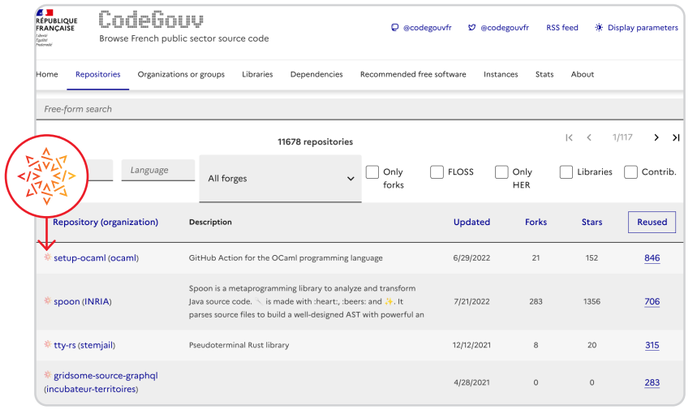
"THESE UNIQUE FEATURES OF SOFTWARE HERITAGE ENSURE AVAILABILITY, GUARANTEE INTEGRITY AND ENABLE TRACEABILITY OF THE SOURCE CODE OF ALL ARTEFACTS IN THE OPEN SOURCE SOFTWARE SUPPLY CHAIN".

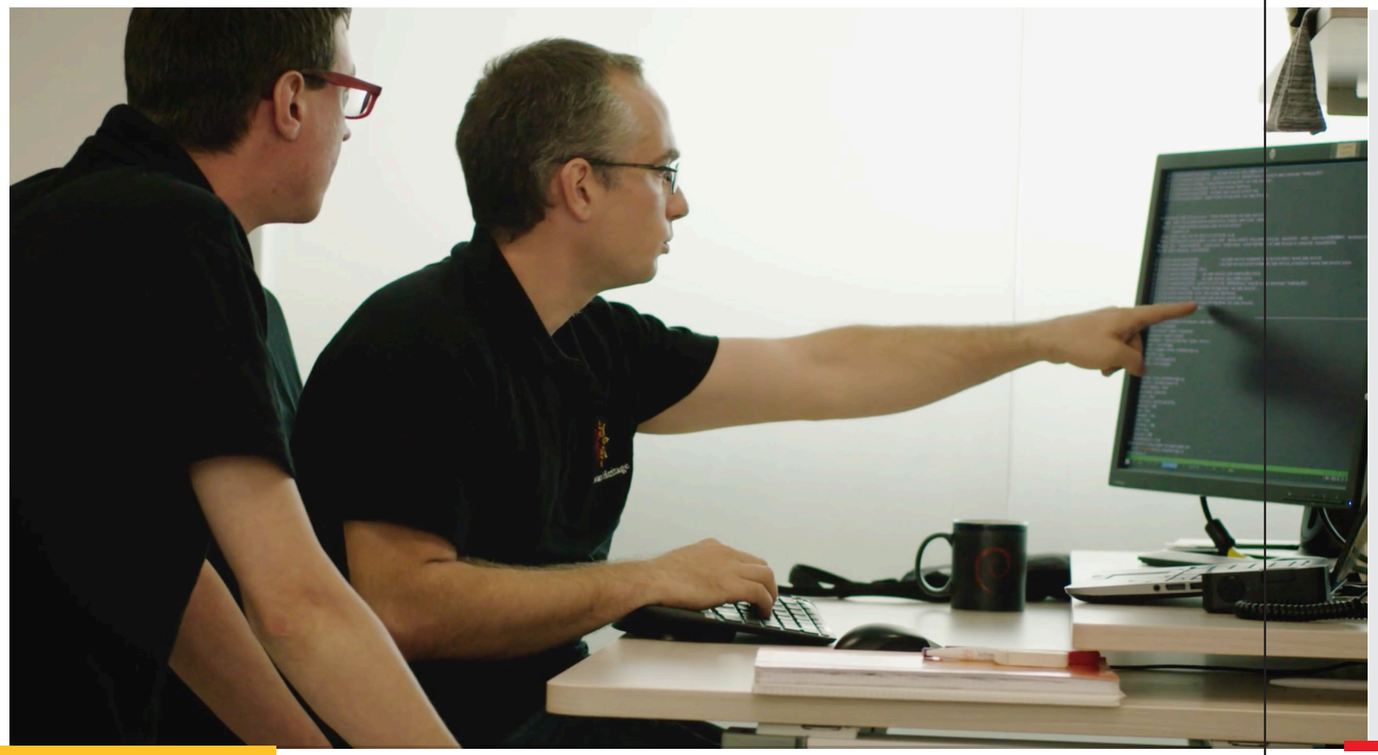
Public Administration

Promoting the sharing of open source solutions created or used by administrations within the European Union [...] results in enhanced collaboration between public administrations

Strasbourg Declaration,
May 2022, European Union

Public administrations strive to make their action transparent to the citizens, and improving the services they provide by sharing and reusing their software.
Software Heritage provides the one-stop archive where all public software can be deposited and referenced, open to all, with the guarantee that it will not disappear.
Software Heritage is now being used by the Open Source mission in the French DINUM to systematically archive the open source software of the french public





Collaboration and community

Alone we go faster, together we go further.
African saying

A broad community is key for succeeding in the long-term mission undertaken by Software Heritage. This is why we are partnering with private funders around the world to provide grants for experts that are willing to engage with the long-term mission of Software Heritage.

Alfred P. Sloan Foundation

A grant from Alfred P. Sloan Foundation has been awarded to Software Heritage specifically to foster the emergence of a community of expert contributors to increase the coverage of the Software Heritage archive. **Seven subgrants** have been distributed, resulting in over 300.000 new repositories being archived.

NGI Zero

Four cascading grants from the NLNet Foundation funded work that allowed Software Heritage to save 250.000 endangered Bitbucket repositories, improve its Mercurial loader, get connectors with Nix and Guix, and experiment with the IPFS distributed file system.

You can help

The Software Heritage archive will serve the needs of the many, from cultural institutions to scientists and industries. Everyone can help us achieving these ambitious goals and there are several ways to help.

Become a sponsor

Pursuing our roadmap for the archive requires significant resources. We welcome companies, institutions, and individuals who would like to join our sponsorship program and sustain the Software Heritage project.

Tackle scientific challenges

Building, maintaining, and exploiting the universal source code archive poses relevant scientific challenges. We welcome scientists who would like to contribute to this mission by participating in our research activities.

Code with us

All the software we develop ourselves is open source. We welcome contributors that are willing to delve into it and help us building the many components that are needed to make the archive progress towards the next milestones.

Host a mirror

Institutions and companies from all around the world are welcome to join our mirror program. This is essential to prevent information loss, and will greatly simplify access to humankind's software heritage.

Meet our team

BEHIND SOFTWARE HERITAGE YOU FIND A TEAM OF PASSIONATE PEOPLE THAT DEDICATE ALL THEIR ENERGY TO THE LONG TERM MISSION OF COLLECTING, PRESERVING AND SHARING THE SOURCE CODE OF ALL PUBLICLY AVAILABLE SOFTWARE. HERE YOU FIND THE COMPOSITION OF THE TEAM AS OF DECEMBER 2022.

Executives

Roberto Di Cosmo *(Founder, CEO)*
Stefano Zacchiroli *(Founder, CTO)*

Advisors

Gérard Berry *(French Academy of Science)*
Jean-François Abramatic *(EIT)*
Julia Lawall *(Inria)*
Serge Abiteboul *(French Academy of Science)*

Project manager

Benoît Chauvet

Open science community manager

Sabrina Granger

Engineers

Jérémy Bobbio
David Douard
Valentin Lorentz
Nicolas Dandrimont
Antoine R. Dumont
Vincent Sellier
Morane Gruenpeter
Antoine Lambert
Jayesh Velayudhan

Visiting scientists

Elisabetta Mori

Communication

Marla da Silva

Ambassadors

At **Software Heritage**, we embarked on a long-term mission to build an open, shared, non-profit, multistakeholder infrastructure by collecting, preserving, and making readily accessible in the Software Heritage Archive the source code of all the software that is publicly available. We know that this is a humbling undertaking, and we will only succeed with the help of a broad community.

To foster community engagement, and accelerate the adoption of Software Heritage in the many fields where it brings groundbreaking benefits, a dedicated ambassador program has been established.


Alexis Lebis


Anna-Lena Lamprecht


Borut Kumperscak


Bostjan Spetic


Bruno Khélifi


Cécile Arènes


Dare Pejić


Flavia Marzano


Gavin Henry


Gerard Coen


Gilmary Gallon

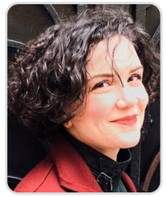

Italo Vignoli


Jaime Arias

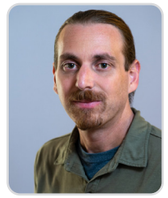

Joenio Marques Da Costa


Julien Caugant


Malin Sandström


Maria-Chiara Prodi


Mohammad Akhlaghi


Neal Fultz


Océane Valencia


Pierre Poulain


Sandrine Layrisse


Vicky Rampin

Becoming an Ambassador

info@softwareheritage.org

Interested in becoming a Software Heritage ambassador? Please tell us a bit about yourself and your interest in the mission of Software Heritage.

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

softwareheritage.org
@swheritage
@SwHeritage