Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Building sustainable infrastructures for (research) software

Roberto Di Cosmo

Founder and Director of Software Heritage
Computer Science Full Professor
Inria and Université Paris Cité

March 11, OECD Workshop on research software

Software Heritage : a decade of commitment (2015-2025)

Laying the foundations (2015-2019)

2015
Project launched
Inria initiates Software Heritage

Jun 30th 2016
Public unveiling
Universal source code archive vision

Apr 3rd 2017
UNESCO
Agreement: Source Code is digital heritage

June 7th 2018
Archive opens
Public access unveiled at UNESCO HQ

Feb 2019
Paris Call
Global policy call on source code

Building *infrastructure* and raising awareness (2019-2023)

Dec 2023
First mirror
Independent full archive online at ENEA

Feb 2022
National roadmap
SWH labeled as a research infrastructure

Jul 2021
Open Science Plan
France recommends and supports SWH for research

Dec 2020
EOSC recognition
Key infrastructure for Research software

Building solutions and strategic initiatives (2023-2025)

Feb 2024
SWH Sec(urity)
Cybersecurity Research using the archive

Nov 2024
CodeCommons
Reference data source for transparent responsible AI

Apr 2025
ISO SWHID
Software Hash Identifier Is ISO/IEC 18670 standard

Jun 2025
Advisory Board
Prepare the organization for the future

Inria

unesco

IBM · Microsoft · TII Technology innovation institute · HUAWEI · cnrs · Infosys · cea · aws · Google · Université Paris Cité · openinventionnetwork

AdaCore · GitHub · Université de Pise · Sorbonne Université · Université de Lorraine · Sant'Anna · NWO · université Paris-Saclay

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Inria with unesco

The largest open source code archive

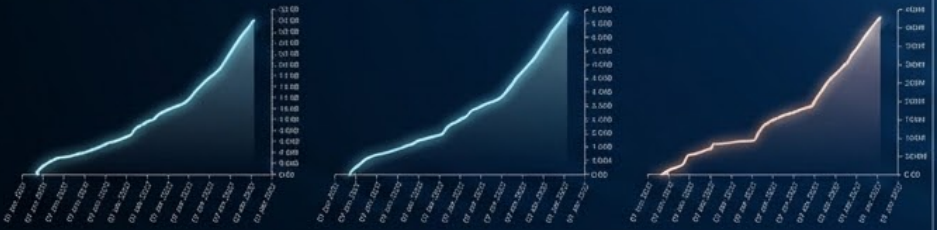
Unique digital common good built in France since 2015

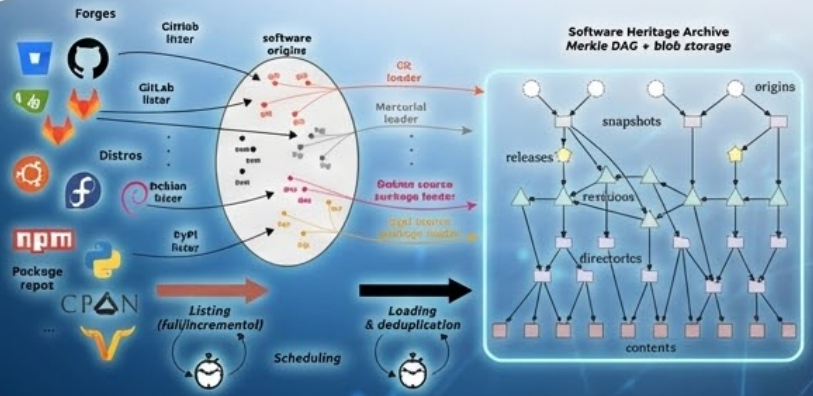
Cultural Heritage | Industry | Research | Public Administration

Source files
28,152,651,759

Commits
5,920,787,012

Projects
429,963,770





5000+ platforms

All versions, all history development in a single graph

- **50 × 10^9** nodes
- **800 × 10^9** edges
- **~ 2 PB** of storage

ensures **availability**
guarantees **integrity**
allows **traceability**

} of source codes

One infrastructure for **multiple** needs
replicated, open, **multistakeholders**

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

The mirror network:

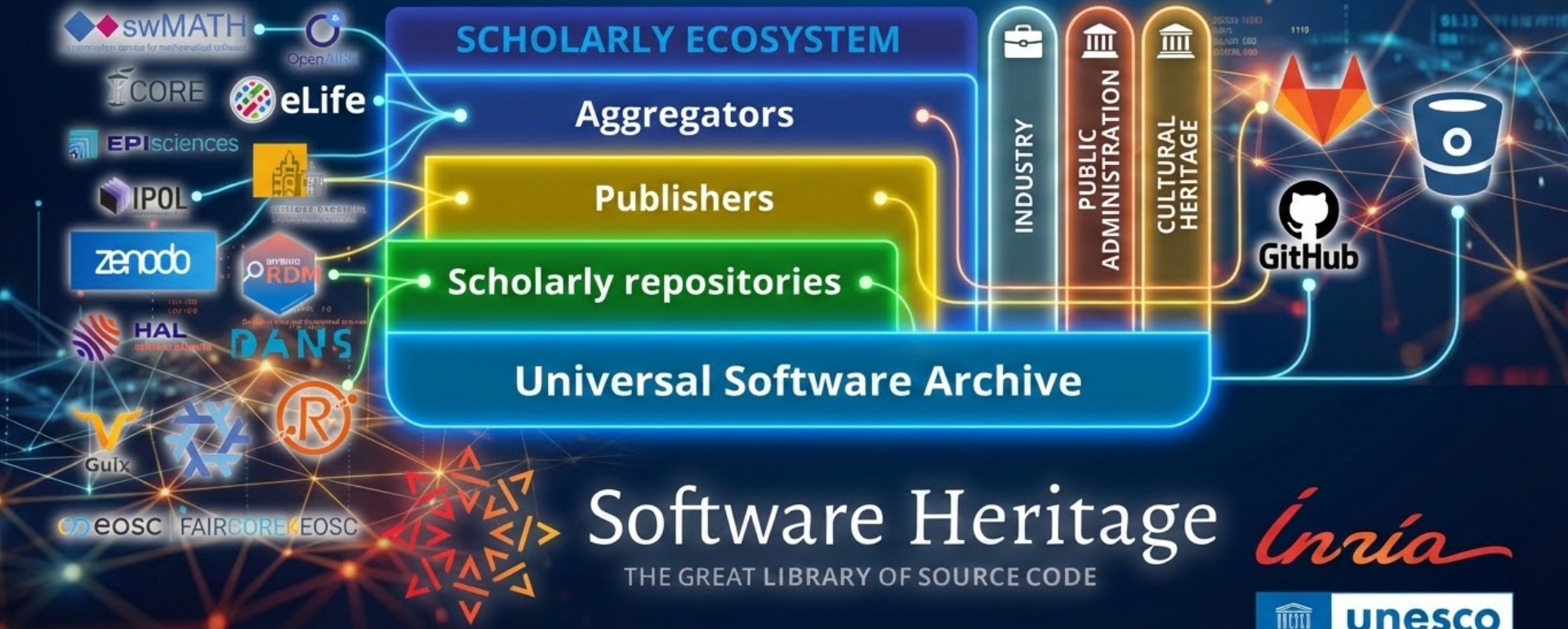
Creates a space for distributed ownership and collaboration

coherence through centralization

resilience through replication

Serving the scholarly ecosystem

Creating connections and mutualizing efforts



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Inria

unesco

From the **Software Sky** to **Institutional Insight**: Anatomy of an Observational Pipeline

Observation parameters (config file)

Stage 0: The Software Sky

The process begins with a complete map of the software universe, built from the Software Heritage archive, containing all observable software 'stars'. This stage is configured by observation parameters and a spectral filter definition (email domain selection).

Phase 1: Target Selection

Identify the stars associated with a specific institution by selecting contributors using institutional identifiers, which isolates the relevant region of the software sky for focused observation.

Quality Gate: Signal Strength Check

Before proceeding, the ploeline verifies that the selected region contains enough observable signals (software contributions) to produce statistically meaningful results.

Eigeat to Nous Raba

Phase 2: Gravitational Cleaning (Deforking)

The process renvents gravitational artifects such as mitrers and fures from the data to retain only the physically meaningful celestial objects (original software projects).

Phase 4: Spectral Enrichment

Each observed object is euriched with spectral data—such as activits, domains, and licenses—collected from multiple observation instruments to provide a deeper understanding.

Adeptive Observation Scheduling

This fecture ensures the automatic relation of observation tokens to bypass iecsument (PPO limits, ensuring continuous data collection.

Observation Cache

A oer vivleen observation log is maintained, preventing redandant measurements and speeding up subsequent analyzes.

Activity

Domains

Licenses

Phase 5: Astrophysical Analysis

The enticed observations are analysed to compute metrics, classily objects into scientific domains, and generate a synthetic, comprehensive view of the observed system.

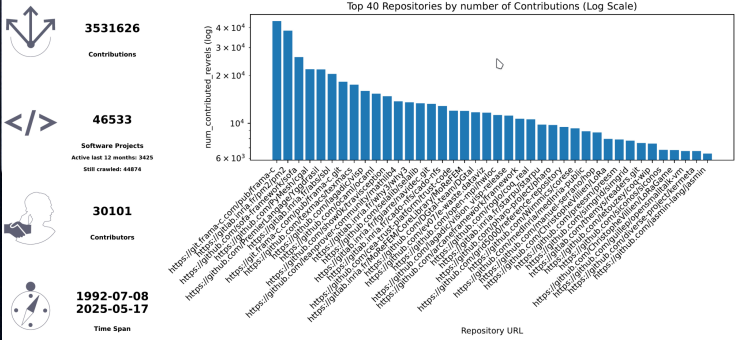
Phase 6: High-Resolution Imaging & Report

3531626
Contributions

46533
Software Projects
Active last 12 months: 3425
Still crawled: 44874

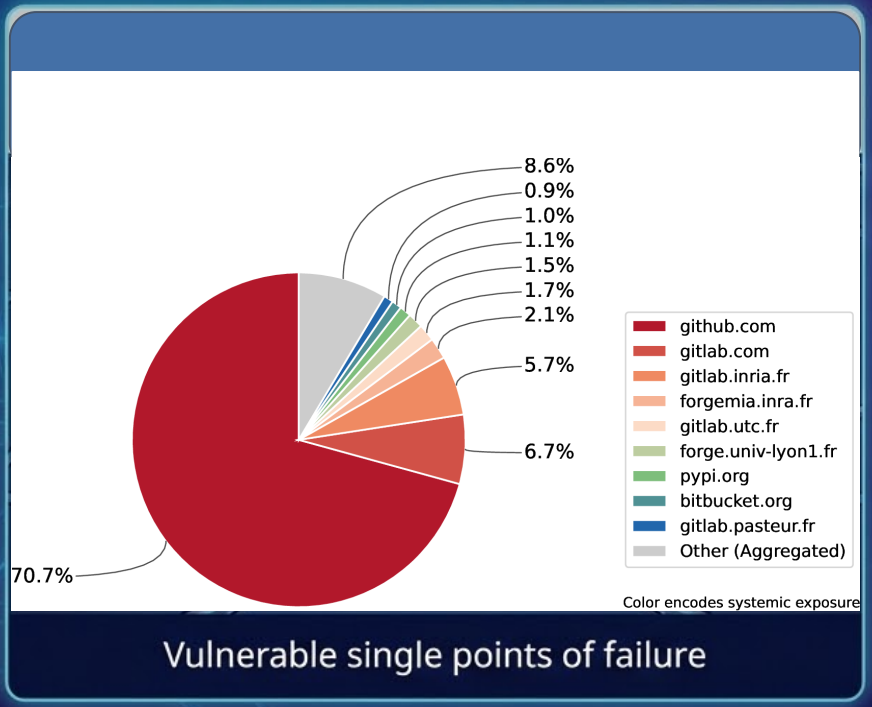
30101
Contributors

1992-07-08
2025-05-17
Time Span



The new frontier: ecumenical resilience

Global Concentration = Global Risk



Vulnerable single points of failure

The systemic response

Today

All the code, also niche and legacy

SWHID, ISO 18670 for integrity/traceability

Enables long term reproducibility

Tomorrow (needs 20x resources!)

strategic autonomy for all

automatic recovery

global fault tolerance

Key principle: avoid balkanisation, mutualize core infrastructure