

From Code to Sovereignty

The Strategic Value of Software Heritage

Roberto Di Cosmo

Director, Software Heritage
Inria and Université Paris Cité

25 August 2026
Fórum RNP 2026 — Brasília



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog



find and reference all
software source code

Universal archive



preserve and share all
software source code

Research infrastructure



enable analysis of all
software source code

A universal archive, as a shared infrastructure

One infrastructure
open and shared

Cultural Heritage



Industry



Research



Public Administration



Inria

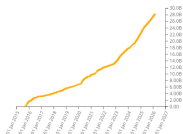
Software Heritage



The largest archive ever built

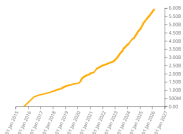
Source files

28,178,230,711



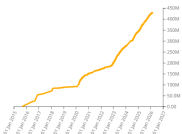
Commits

5,925,711,720



Projects

430,218,860



Directories

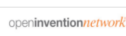
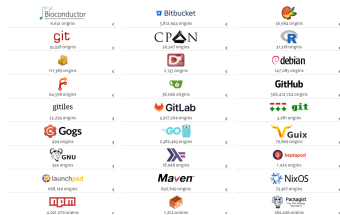
22,013,760,256

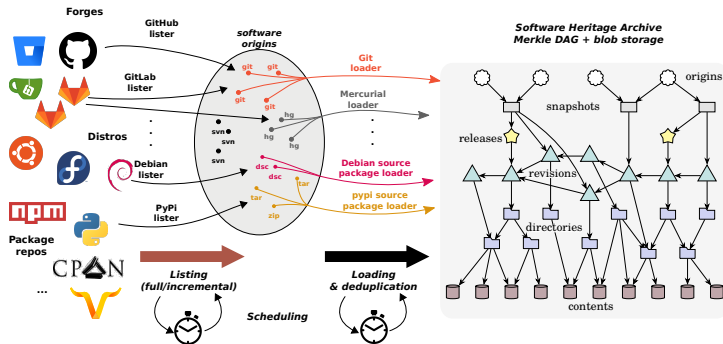
Authors

105,360,419

Releases

147,686,240



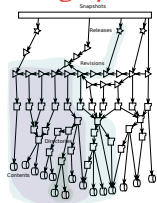


Global development history permanently archived in a uniform data model

- over 29 billion unique source files from over 438 million software projects
- ~3PB (compressed) blobs, ~60 B nodes, ~1 trillion edges

A revolutionary infrastructure

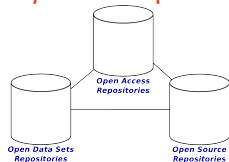
The *graph* of public software development



All software development
in a **single graph** ...

- enable traceability

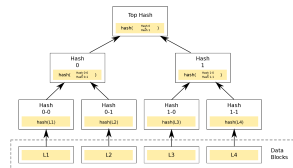
A pillar of Open Science



Reference **archive** of
Research Software

- reproducibility
- reference

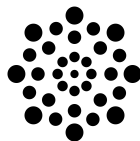
The *global ledger* of public code



... a **Merkle** graph

- ensure integrity

Reference platform for *Big Code* **uniform** data structure



- large scale studies
- cybersecurity
- machine learning, AI, ...

Groundbreaking services

Data for Transparent AI

——</> </> Traceable source code

</> </> </> for LLMs

</> </>—— CodeCommons



CyberSecurity



Make open-source
more secure



Support Open Source Program Offices



OSPO
RADAR

Track, find and reuse
research software



Strategic Metrics and Insights The Public Code Observatory with UN and UNESCO



Software Heritage

Strategic Software
Insights Report



4 520 research articles cite Software Heritage (*Google Scholar, 2025*) • 2 M+ Save Code Now deposits

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

Observation parameters (config file)

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 pipeline verifies that the selected region contains enough observable signals (software contributions) to produce statistically meaningful results.

Phase 2: Gravitational Cleaning (Deforking)

The process removes gravitational artifacts such as mirrors and forks from the data to retain only the physically meaningful celestial objects (original software projects).

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 4: Spectral Enrichment

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

Adaptive Observation Scheduling

This feature ensures the automatic relation of observation tokens to bypass instrument (DPO) limits, ensuring continuous data collection.

Observation Cache

A per-widened observation log is maintained, preventing redundant measurements and speeding up subsequent analyses.

Phase 5: Astrophysical Analysis

The enriched observations are analysed to compute metrics, classify 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: 9029
BRO covered: 44819



30101

Contributors

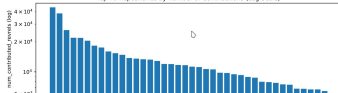


1992-07-08

2025-05-17

Time Span

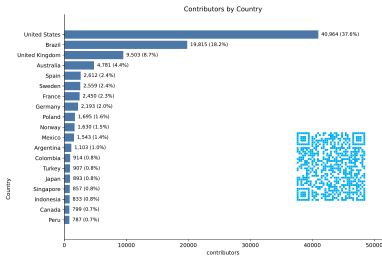
Top 40 Repositories by number of Contributions (Log Scale)



Repository URL

Second in the world

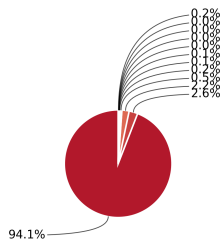
government code



Public Code Observatory — UN, UNESCO, SWH
Brazil 2nd of 193 UN member states, July 2026

94 % of it lives on GitHub

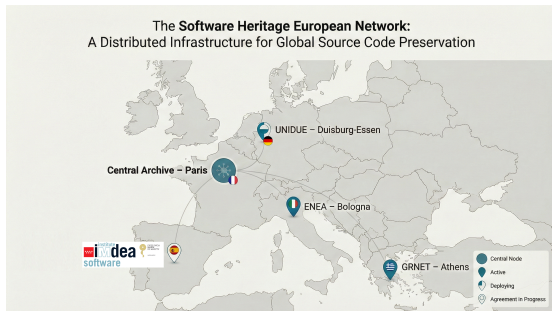
all of .br



All Brazilian open source — not just government or academic: **256 908** projects, every .br domain.

A top-two producer of public code — on one platform, in one jurisdiction.
Global concentration = global risk.

The mirror network today



Four mirrors, all in Europe. Brazil would be the first beyond.

Soberania sem rivalidade. Controle sem exclusividade.

Two decisions, one moment

- 1 **Fund the mirror at LNCC**
continuity of world source code, on
Brazilian soil
- 2 **Join the Foundation as a major
sponsor**
join the governance

Each country can host it.

No country controls it.



10 years report