Building the infrastructure Europe needs

Roberto Di Cosmo

Director, Software Heritage
Inria and Université Paris Cité

Launch of the Strategic Research and Innovation Agenda for the Future of Software
Brussels, 23 September 2026 — online from Cork



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Inria with unesco

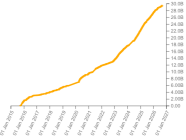


The largest open source code archive: one infrastructure, open, shared, non profit

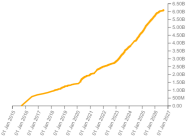
Unique digital common good *built in France since 2015*

Cultural Heritage | **Industry** | **Research** | **Public Administration**

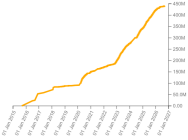
Source files
29,523,977,803

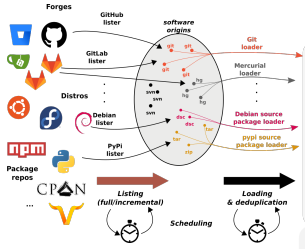


Commits
6,112,189,339



Projects
441,196,029





5000+ platforms

All versions, all history development in a single graph

- **58.7 × 10⁹** nodes
- **1041 × 10⁹** edges
- ~ **3 PB** of storage

A revolutionary **infrastructure** ensures **availability** guarantees **integrity** enables **traceability**

Microsoft · TII Technology Innovation Institute · HUAWEI · cnrs · Infosys · cea · Ministère de l'Enseignement supérieur et de la Recherche · aws · Google · Université Paris Cité · openinventionnetwork

AdaCore · Canada Research Knowledge Network · GitHub · Premier Ministre · Università di Pisa · Sorbonne Université · Scuola Normale Superiore · Université de Lorraine · Sant'Anna · NWO · université PARIS-SACLAY

The SRIA already names what Europe is missing

The agenda's own infrastructure list

"Dedicated infrastructure funding for shared and sovereign European software capabilities such as

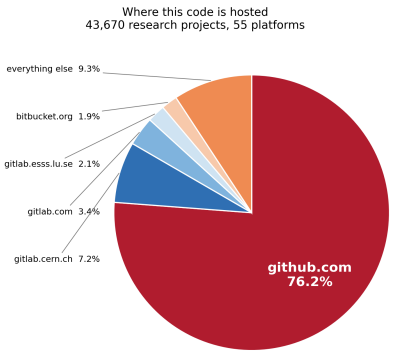
Federated software forges	Dataset archives for AI on code	Benchmarks	Reproducible build clusters	Testing and verification
to be built	under way	to be built	to be built	to be built

Every one of these rests on the same thing: the source code has to still be there, and be identifiable.

That layer is not a proposal. It exists, it is European, and it took ten years to build.

Software research now has its own infrastructure.

What we measured, and it is about your code



The public code of the institutions behind Europe's research infrastructures: **43,670** projects, each one classified as research, **24,457** contributors.

76% on **one platform**, under one jurisdiction, against 7.2% on CERN's own forge and 2.1% on the European Spallation Source's

5.4% **already gone** from where they were published, alive only because they were archived

13 September 2026: two hours of global outage on that platform, from an internal maintenance job. **March 2024**: `xz-utils` unreachable for eleven days, readable throughout in the archive.

Software Heritage archive, June 2026. Projects with $\geq$ 25 contributions from the 196 institutional domains of the five EOSC Science Clusters, each classified as research or not.

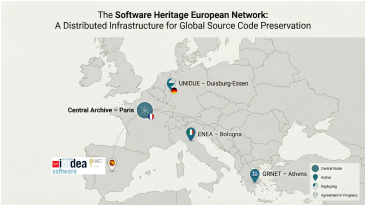
Infrastructure is what other things are built on

Load-bearing today



Since 2019 **GNU Guix** rebuilds from source and **falls back on the archive automatically** when a source has vanished. **Nix** lists its sources every day; every Nix hash maps to a SWHID.

Ecumenical digital sovereignty



One non profit open international core infrastructure. Each **mirror** run by an *independent organisation, in its own jurisdiction, on its own stack*: not one fortress, plural custody.

Plus the identifier layer that makes any of this checkable: **SWHID, ISO/IEC 18670**, intrinsic identifiers for whole source trees.

What Europe still has to build, and who owns it

The missing pieces

- federated forges, so that hosting is European and plural, with archiving by design rather than by goodwill
- reproducible build clusters, the operational test of whether sovereignty is real
- corpora for AI on code with provenance, licences and traceability
- benchmarks, testing and verification that the whole field can share

France did it in one document

The French Open Science Strategy 2026–2033 asks, in the same objective, for a **sovereign development forge** *and* for archiving on Software Heritage, and names supporting the archive as an action of the national plan.



Software needs an **institutional home** in Europe, and this agenda needs a **sustained programme** behind it.

Infrastructure is not a project: it is what projects are allowed to depend on.

The question is no longer whether Europe can build this, but whether it will own what it has already built.