Databases at risk

The **software layer** of research digital sovereignty



Roberto Di Cosmo

Director, Software Heritage · Inria and Université Paris Cité @rdicosmo roberto@dicosmo.org

4th ESFRI Stakeholder Forum Meetup

Session 2: Databases at risk and Research Data Sovereignty

Cork, 22 September 2026

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

 · CC-BY 4.0

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Inria with unesco



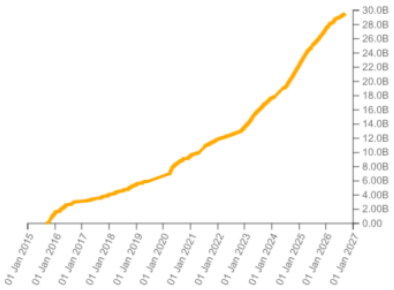
The largest open source code archive: one infrastructure, open, shared, non profit

Unique digital common good *built in France since 2015*

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

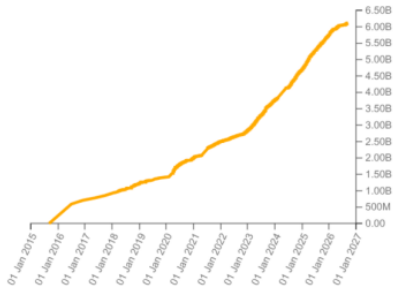
Source files

29,523,977,803



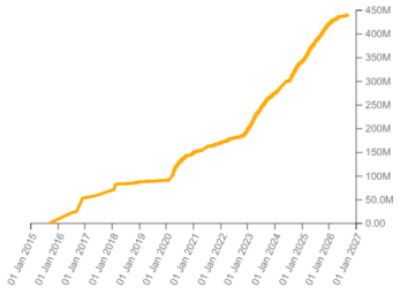
Commits

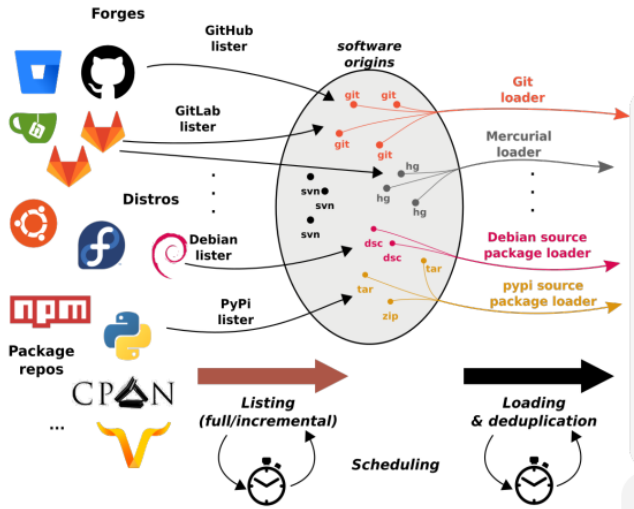
6,112,189,339

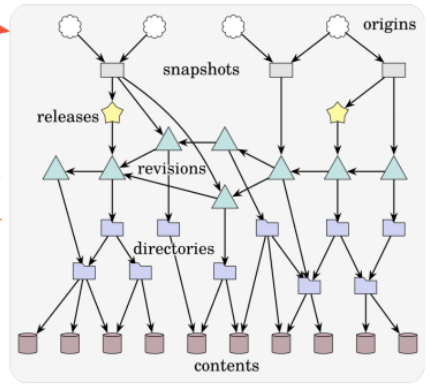


Projects

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 · GitHub · Università di Pisa · Sorbonne Université · Université de Lorraine · Sant'Anna · NWO · université Paris-Saclay

The software of your Science Clusters, measured by Software Heritage

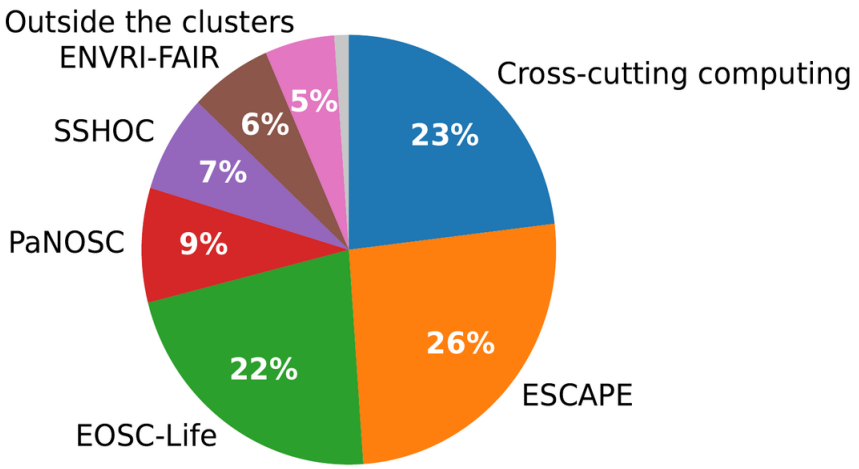
43,670 research code projects, **24,475** contributors, **20** countries.

76% of them on **github.com**, 7.2% on CERN's own forge, 2.1% on ESS's.

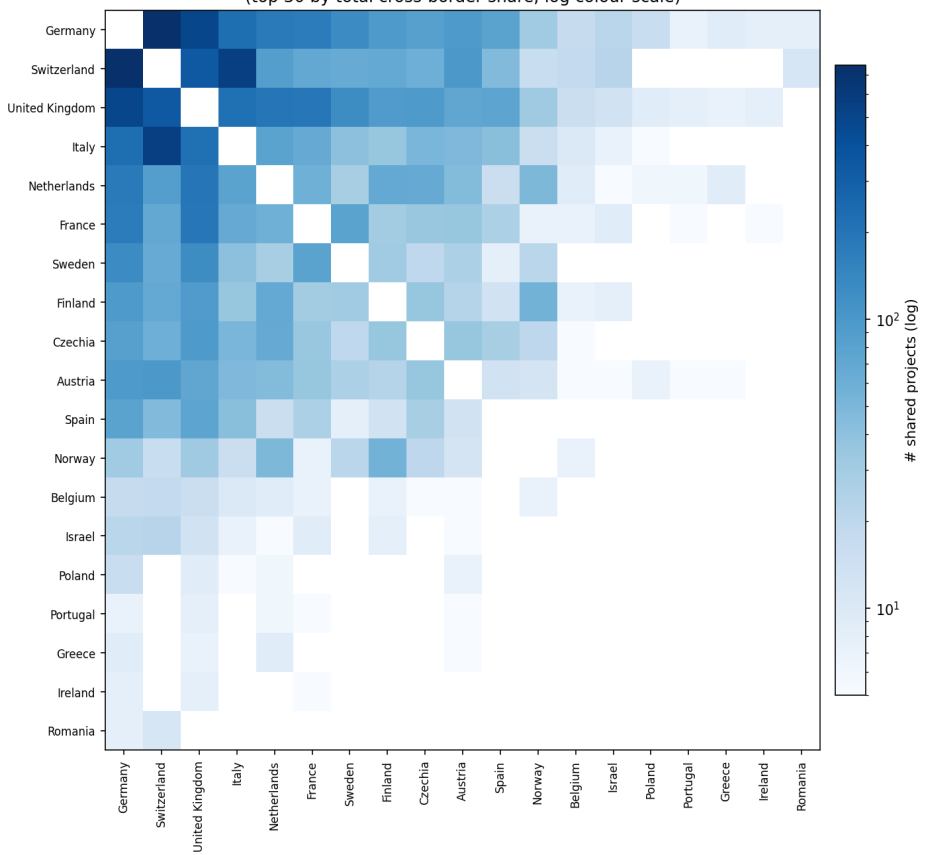
5.4%, **2,356 projects**, disappeared already, saved in Software Heritage.

the full report

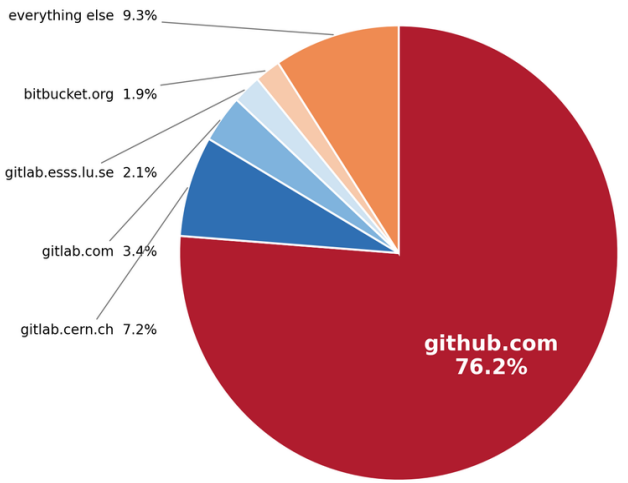
What this research code serves
43,670 classified research projects



Cross-border collaboration: shared projects per country pair
(top 30 by total cross-border share; log colour scale)



Where this code is hosted
43,670 research projects, 55 platforms



Software Heritage

Strategic autonomy: who *controls* your code?

The danger: 76% of your research code on **one platform, one jurisdiction**

Forges are not archives

- **platforms close** (Google Code, Gitorious, … 1M projects ~~lost~~ saved in Software Heritage) **projects change, move, disappear**
- **30 March 2024**: xz-utils accounts suspended, repository unreachable until 10 April (*but available on SWH!*)
- **13 September 2026**: two hours of global GitHub outage

The way forward

Resilience

May 21, 2025

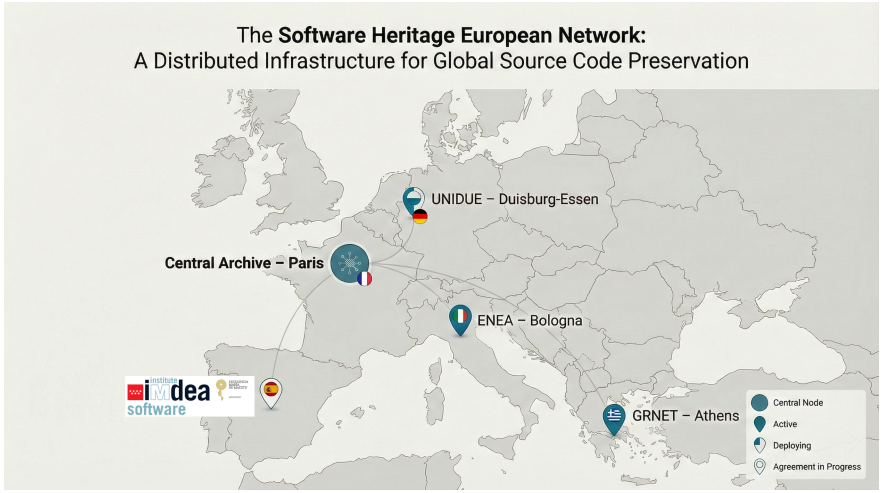
How Software Heritage ensures reliable Guix deployment



Reproducibility in research is a growing challenge. After researchers publish their findings, the underlying software — the source code that drives their results — often disappears or becomes unusable over time. This makes verifying and building upon past work incredibly difficult.

Since 2019 **GNU Guix** falls back to the archive **automatically**. **Nix** is next.

Ecumenical digital sovereignty



One non profit open international core infrastructure.

Multiple **mirrors** run by an *independent organisation, in their own jurisdiction, on their own stack.*

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE

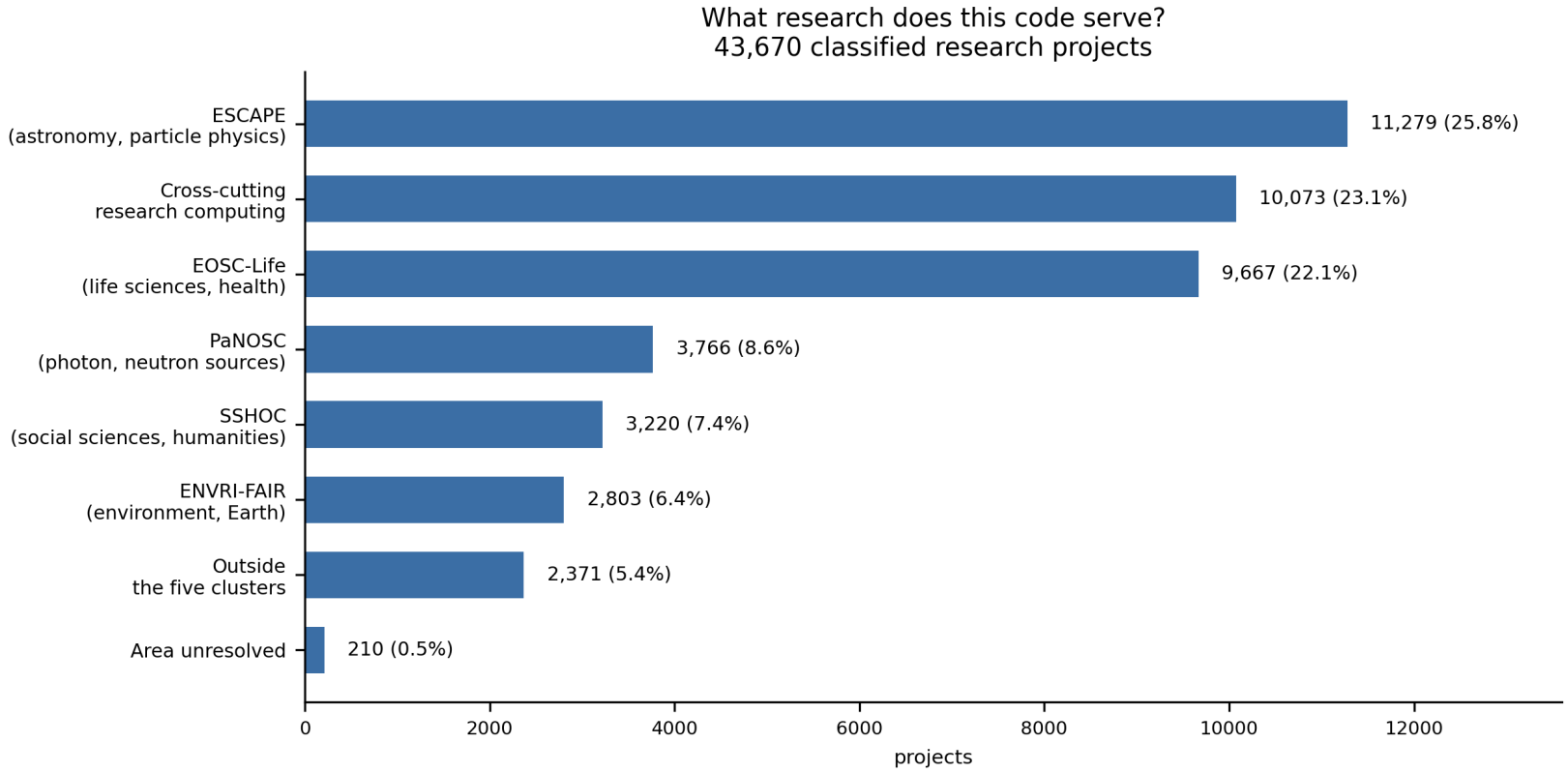
Ouvrir la science !



Multiannual support for Software Heritage as a strategic **research infrastructure**

Appendix

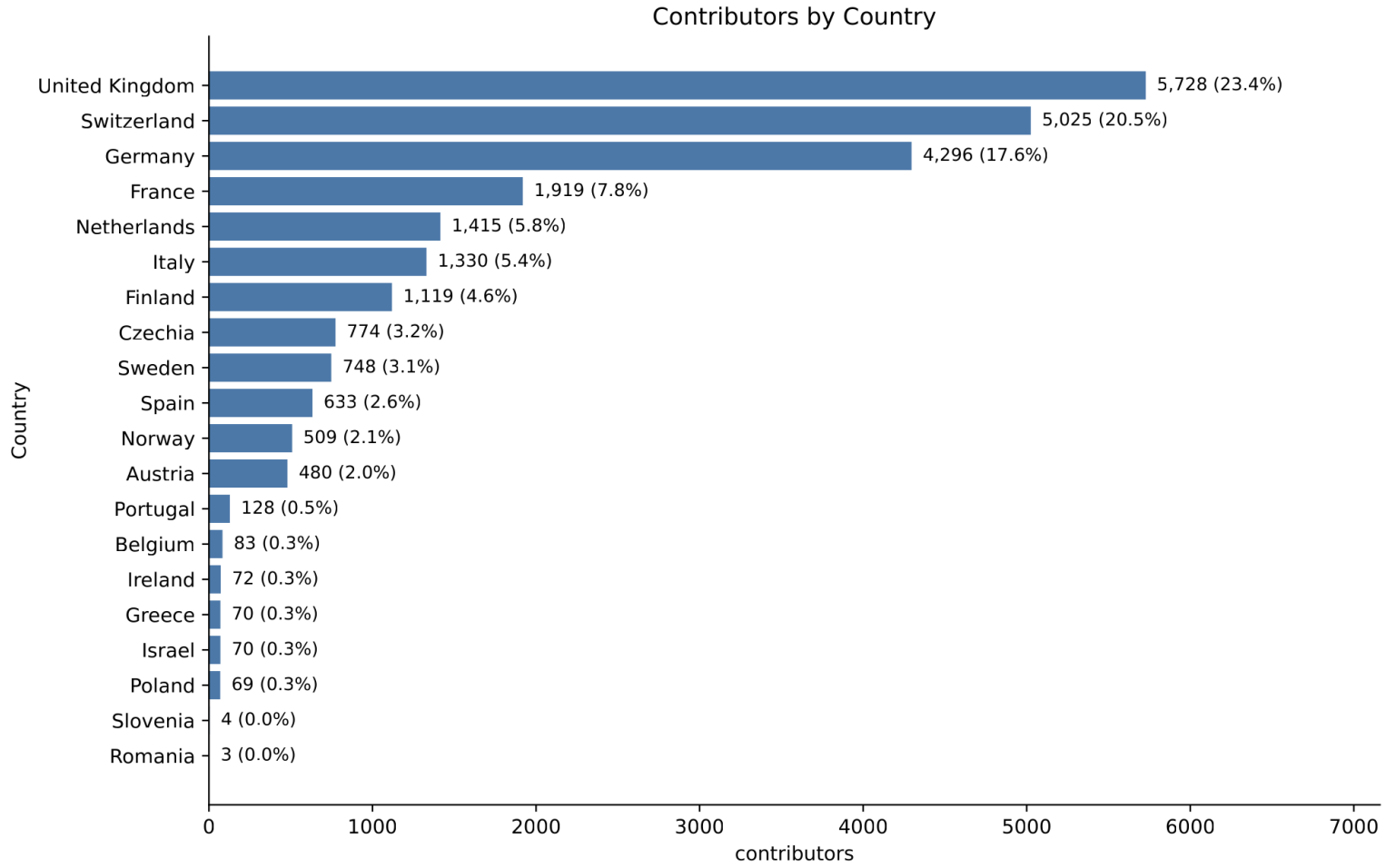
What your Science Clusters actually build



ESFRI Science Clusters, projects with at least 25 institutional contributions. Every project at this floor was judged individually by two models reading repository metadata, contributor email domains and the archived README, with a third model settling disagreements; research data is kept and marked, never dropped. 491 projects (1.1%) carry no cluster label.

43,670 projects and **24,475** contributors from the 196 domains; 1.1% carry no cluster label.

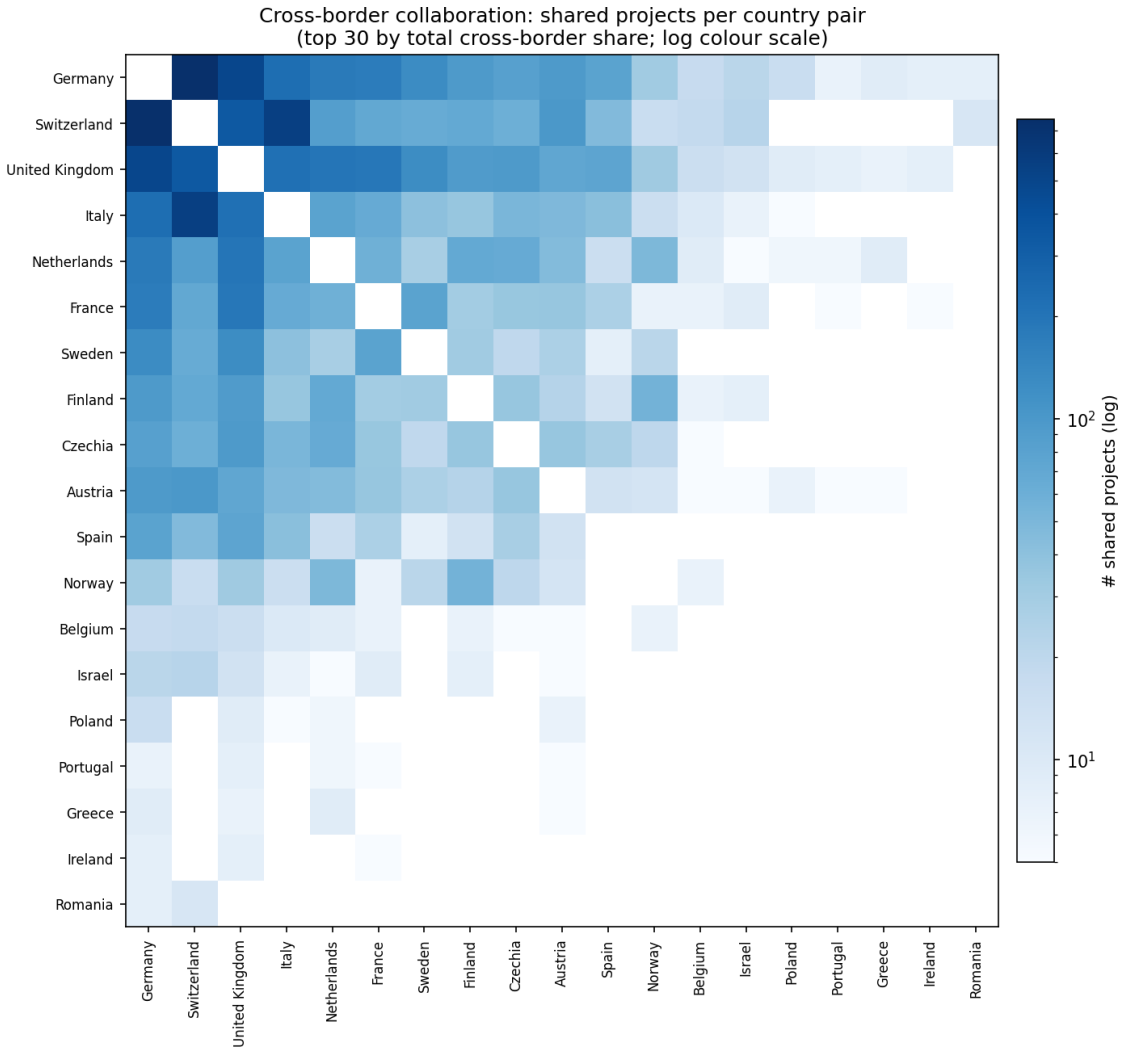
Who writes it: contributors by country



All 24,475 contributors over the 43,670 projects, attributed to the country of their institutional email domain. CERN counts as Switzerland.

The code is already written across borders

Chart



Facts

Projects shared by contributors from two countries of the ESFRI community:

Germany Switzerland	**756**
Italy Switzerland	553
Germany United Kingdom	488
Switzerland United Kingdom	340

The software layer is **already federated** in practice. Its preservation is not.

102 country pairs, 6,430 shared projects, over the 43,670 classified research projects whose contributors resolve to a country.

Research infrastructures' own forges: what the archive holds today

1.25 self-hosted forge	research infrastructure	origins archived
code.ill.fr	Institut Laue-Langevin	189
gitlab.eso.org	ESO / ELT	1
gitlab.et-gw.eu	Einstein Telescope	0

Across the 63 RIs of the ESFRI Roadmap 2021: **44 of 49** forge entries are on github.com

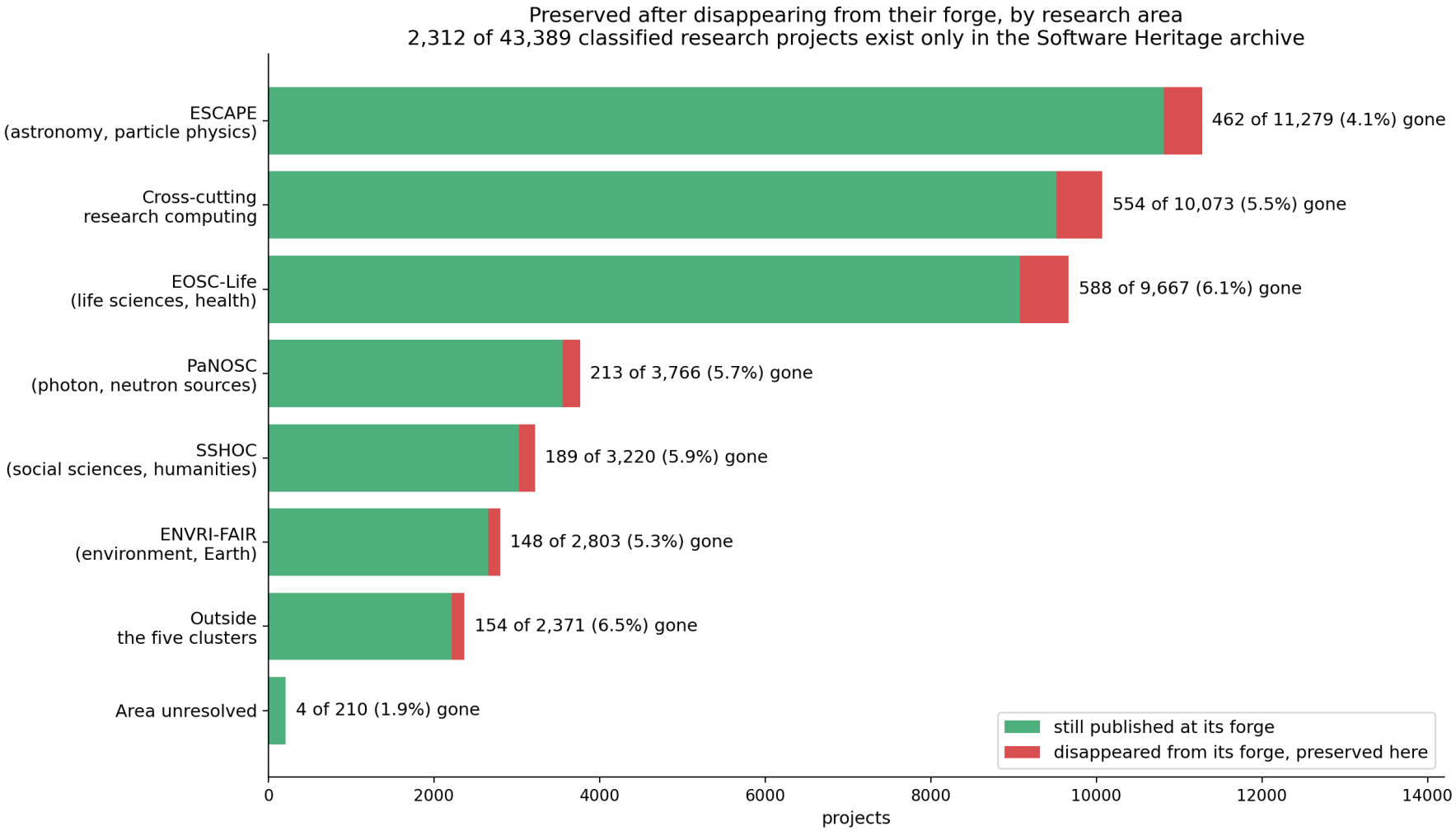
Self-hosting gives **control**, not **durability**: sovereignty needs both.

Tell us where your forge is, and we will archive it.

Lower bounds over what Software Heritage has indexed: a zero means the archive holds nothing from that forge yet, not that the code is lost.

Software Heritage

Which research areas lose most



Best held: astronomy and particle physics (ESCAPE), 4.1% gone. Largest absolute loss: life sciences and health (EOSC-Life), 588 projects.

Method, in one slide

- **Perimeter**: 196 institutional email domains of the participants of the five EOSC Science Clusters, from CORDIS, verified by DNS and MX
- **Selection**: every public project in the archive with at least 25 revisions or releases authored from those domains
- **Classification**: every one of the 122,517 projects above the threshold judged individually by two language models reading its metadata and its archived README, a third model settling disagreements, every action decided by code and not by the model. 71,651 dropped as non-research; of the 50,866 that remained, 43,670 have a contributor at an institution inside the ESFRI perimeter
- **No longer found**: projects the archive no longer reaches at their original URL, still preserved in it
- **Limits**: author email is a **floor**: contributions from personal addresses are missed, and institutions whose staff write from other domains are under-counted

 · CC-BY 4.0

Software Heritage