No Knowledge without Source

Collecting, Preserving and Sharing Software in a Risky World

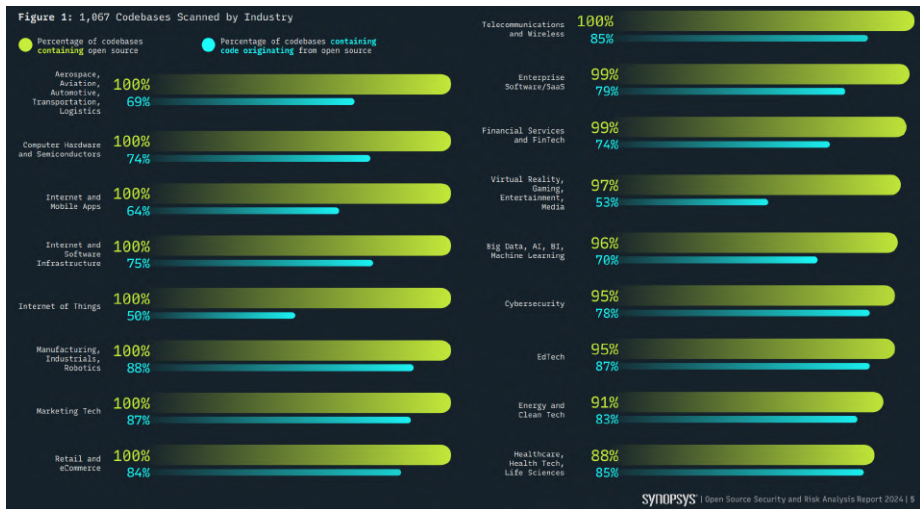**Roberto Di Cosmo**

Inria and Université Paris Cité

Director, Software Heritage

Co-chair, Software College,

French Open Science Committee

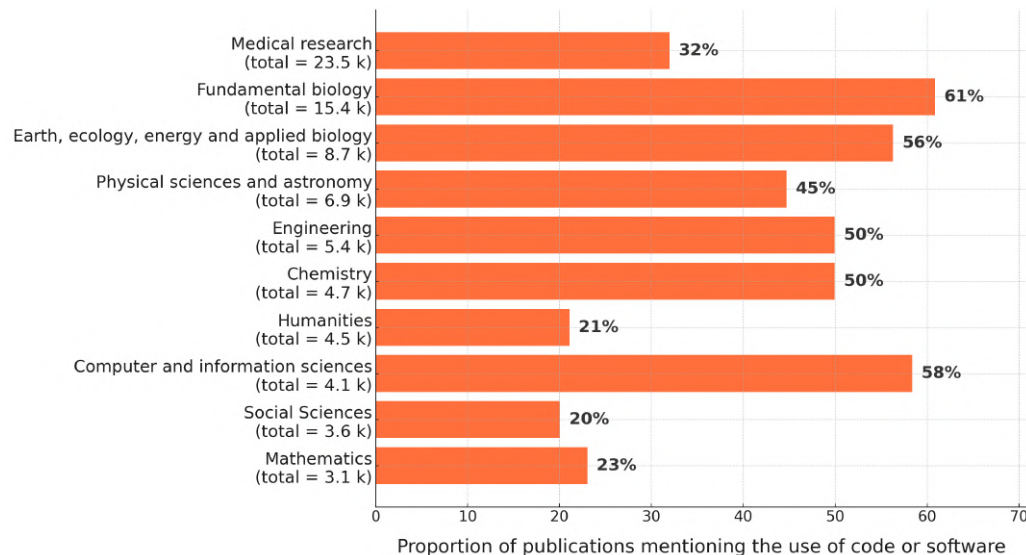https://dicosmo.org   @rdicosmo

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

1

Open Source Software: "~~data~~ altruism" everywhere

Industry

Open Source Security/Risk Analysis 2024

- OSS in 96--97% audited commercial codebases
- ~77% of code within those codebases is OSS
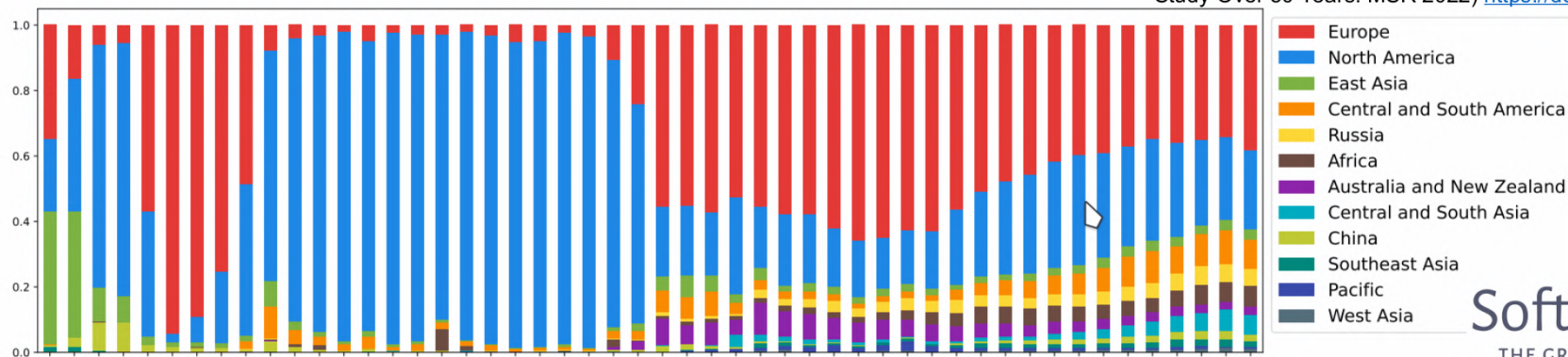- Avg. ~900+ OSS components/app

Academia

French Open Science Monitor 2025
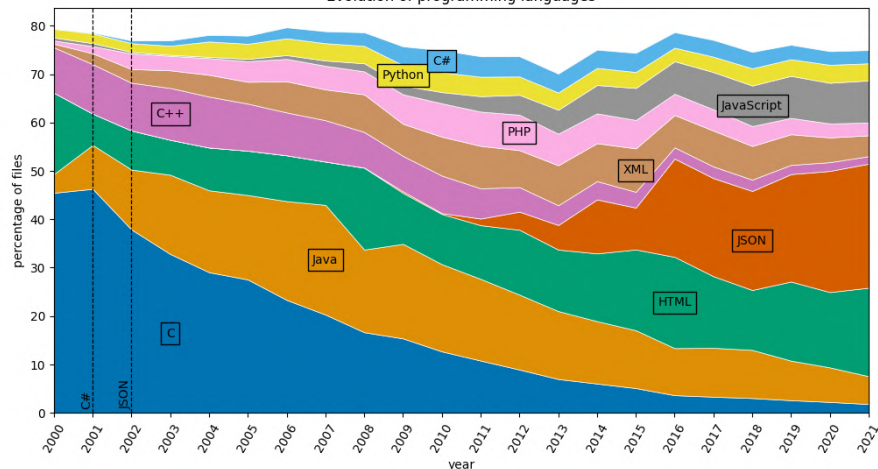
Open Source Software: global and growing

Ratio of commits by world zone over the 1971–2020 period.

Davide Rossi and Stefano Zacchiroli Geographic Diversity in Public Code Contributions: An Exploratory Large-Scale Study Over 50 Years. MSR 2022) https://doi.org/10.1145/3524842.3528471
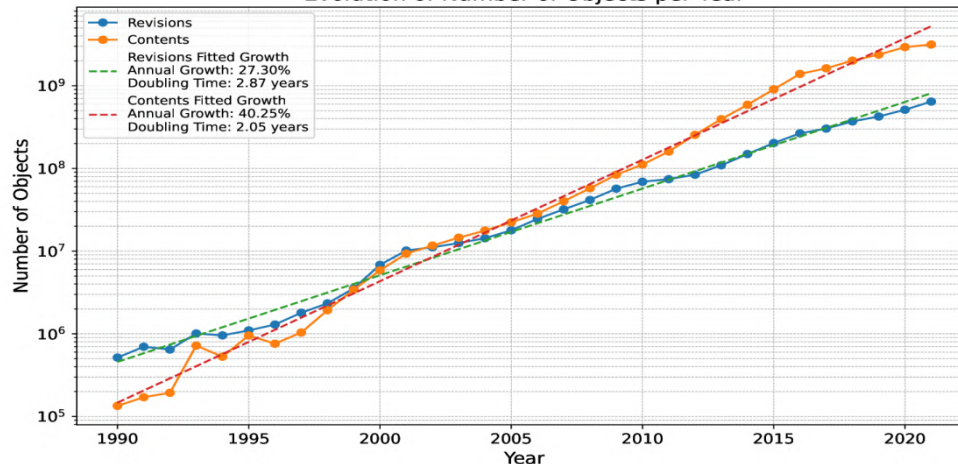
Legend:
- Europe
- North America
- East Asia
- Central and South America
- Russia
- Africa
- Australia and New Zealand
- Central and South Asia
- China
- Southeast Asia
- Pacific
- West Asia

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Evolution of programming languages

Evolution of Number of Objects per Year

- Revisions
- Contents
- Revisions Fitted Growth
  Annual Growth: 27.30%
  Doubling Time: 2.87 years
- Contents Fitted Growth
  Annual Growth: 40.25%
  Doubling Time: 2.05 years

Adèle Desmazières, Roberto Di Cosmo, Valentin Lorentz

50 Years of Programming Language Evolution through the Software Heritage looking glass
In: IEEE, (Ed.): Mining Software Repositories, Ottawa (Canada), Canada, 2025.

It is a risky (digital) world

**Digital fragility
(one root of non reproducibility)**

2015: the first big bad news

Google Code and Gitorious.org shutdown: ~1M endangered repositories

- broken links in the web of knowledge (my papers too)

Big bad news keep coming in

- summer 2019: BitBucket announces Mercurial VCS sunset
- july 2020: BitBucket erases *250.000+* repositories (including research software)
- summer 2022: GitLab.com considers erasing all projects that are inactive for a year

In Academia too!

- 2021: Inria's old gforge is unplugged… breaks the Opam build chain for OCaml

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*

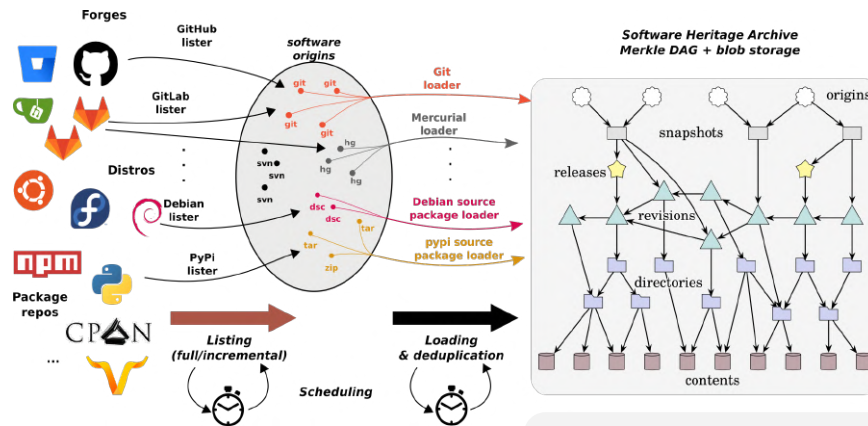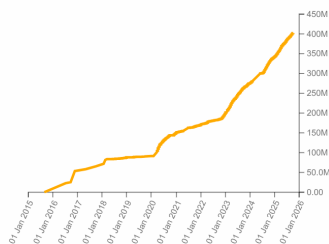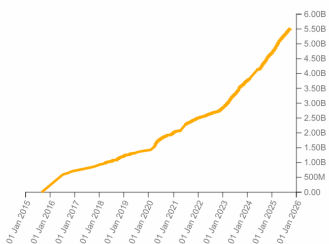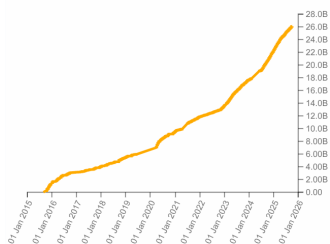| Cultural Heritage	Industry	Research	Public Administration

Source files
26,098,911,720

Commits
5,511,094,786

Projects
402,119,077

Software Heritage Archive
Merkle DAG + blob storage

5000+ platforms

**All versions, all history
development in a single graph**

- **50 × 10⁹** nodes
- **800 × 10⁹** edges
~ **2 PB** of storage

ensures **availability**
guarantees **integrity**
allows **traceability**
} of source
codes

A **unique infrastructure** to preserve the knowledge
embedded in software source code

More risks and challenges

Collateral damage from

- EU Copyright Directive

- GDPR

New obligations

e.g. Cyber resilience act
(ongoing, high impact)

Coping with the (Gen)AI tidal wave

Closed behind APIs

✗ Model weights not available

- *Can't run the model locally*
- *Can't inspect model representation*
- *Limits fine-tuning abilities*
- *Limits user freedom (personal data leakage)*

Open weights

✗ Training data not disclosed

- *Creators don't know if their data is used*
- *There's no way to remove it*
- *Can't inspect data for biases*
- *Potential benchmark contamination*
- *Limits scientific reproducibility*

It's time to focus on key challenges for training data
availability
transparency
integrity

Open source AI definition

Sufficiently detailed information about the data used to train the system so that a skilled person can build a substantially equivalent system.

AI Act

Article 53: special exception for *providers of AI models released under a free and open-source licence[...] and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available.*

PRESS RELEASE | Publication 24 July 2025

Commission presents template for General-Purpose AI model providers to summarise the data used to train their model

What should Software Heritage do?

Founding principles at Software Heritage

Principles

October 19, 2023

Software Heritage Statement on Large Language Models for Code

1. Knowledge derived from the Software Heritage archive must be given back to humanity, rather than monopolized for private gain. The resulting *machine learning models* must be made available under a suitable open license, together with the documentation and toolings needed to use them.

2. The *initial training data extracted from the Software Heritage archive* must be fully and precisely identified by, for example, publishing the corresponding SWHID identifiers (note that, in the context of Software Heritage, public availability of the *initial training data* is a given: anyone can obtain it from the archive). This will enable use cases such as: studying biases (fairness), verifying if a code of interest was present in the training data (transparency), and providing appropriate attribution when generated code bears resemblance to training data (credit), among others.

3. Mechanisms should be established, where possible, for authors to exclude their archived code from the training inputs before model training begins.

Dataset — The Stack v2 — BigCode × Software Heritage

Model — StarCoder2 — built by BigCode — supported by servicenow × 🤗 × nVIDIA
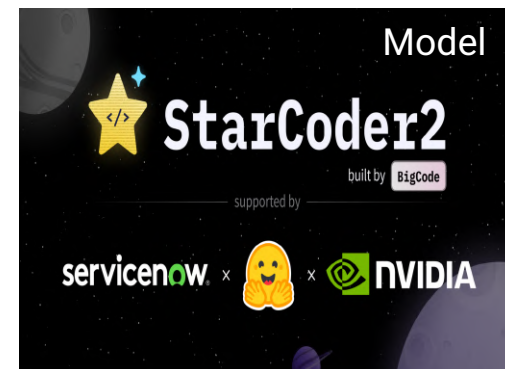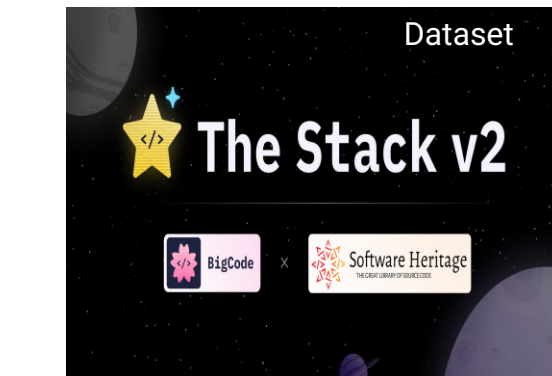
Yes, it's possible

but it's hard to do it well

Lessons learned

Principles

1. Knowledge derived from the Software Heritage archive must be given back to humanity, rather than monopolized for private gain. The resulting *machine learning models* must be made available under a suitable open license, together with the documentation and toolings needed to use them.

2. The *initial training data* extracted from the *Software Heritage archive* must be and precisely identified by, for example, publishing the corresponding SWHID identifiers (note that, in the context of Software Heritage, public availability of the *initial training data* is a given: anyone can obtain it from the archive). This will enable use cases such as: studying biases (fairness), verifying if a code of interest was present in the training data (transparency), and providing appropriate attribution when generated code bears resemblance to training data (credit, among others.

3. Mechanisms should be established, where possible, for authors to exclude their archived code from the training inputs before model training begins.

Transparency is easy:
use SWHID (under ISO/IEC standardisation)
and Software Heritage

(Re)use of the data is tricky: who is *the real owner of a source code file*? What are the real use rights?

- **Building a qualified training set** is expensive (includes **license detection at massive scale**)

- **Attribution** on model output **is challenging**
  (800 billion edges, and counting!)

Software Heritage SCAI *members* can
- **deposit** source codes
- **publish** their SWHID

Software Heritage
Membership Program Source Code & AI

Software Heritage

today!

Need **a coordinated effort** to address these issues

It's time to build a

Code Commons

CODE COMMONS

Availability: Easy access to all relevant data for software (source code, PR, issues, discussions, etc.)
Shared harvesting: we do it only once!

Traceability: Identify and make available the data used for training

Ethics: Provide tools to verify the provenance and attribution of generative AI outputs

Data Sources → Gather data → **Raw Data** → Clean and Process → **Training Data** → Train the model → **Model** → Test the model → **Qualified Model** → Production

Structuring: Organize and connect the various data sources to create a coherent training set.

Efficiency: Facilitate the extraction of qualified datasets to build high-performance models.

Inria · Software Heritage THE GREAT LIBRARY OF SOURCE CODE · cea · TWEAG by Modus Create

Sant'Anna School of Advanced Studies – Pisa · UNIVERSITA DEGLI STUDI DI TORINO · UNIVERSITÀ DI PISA · ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA · AboutCode

Solutions

GENCI HPC at the service of knowledge

bpifrance | SERVIR L'AVENIR

FRANCE 2030

CODE COMMONS : BIRD'S EYE VIEW (technology)

Utilisation de SWH Code Commons

Models training operators

AI Models

Entités utilisatrices

1 - Prompt

2 - Raw answer

3 - Qualified answer
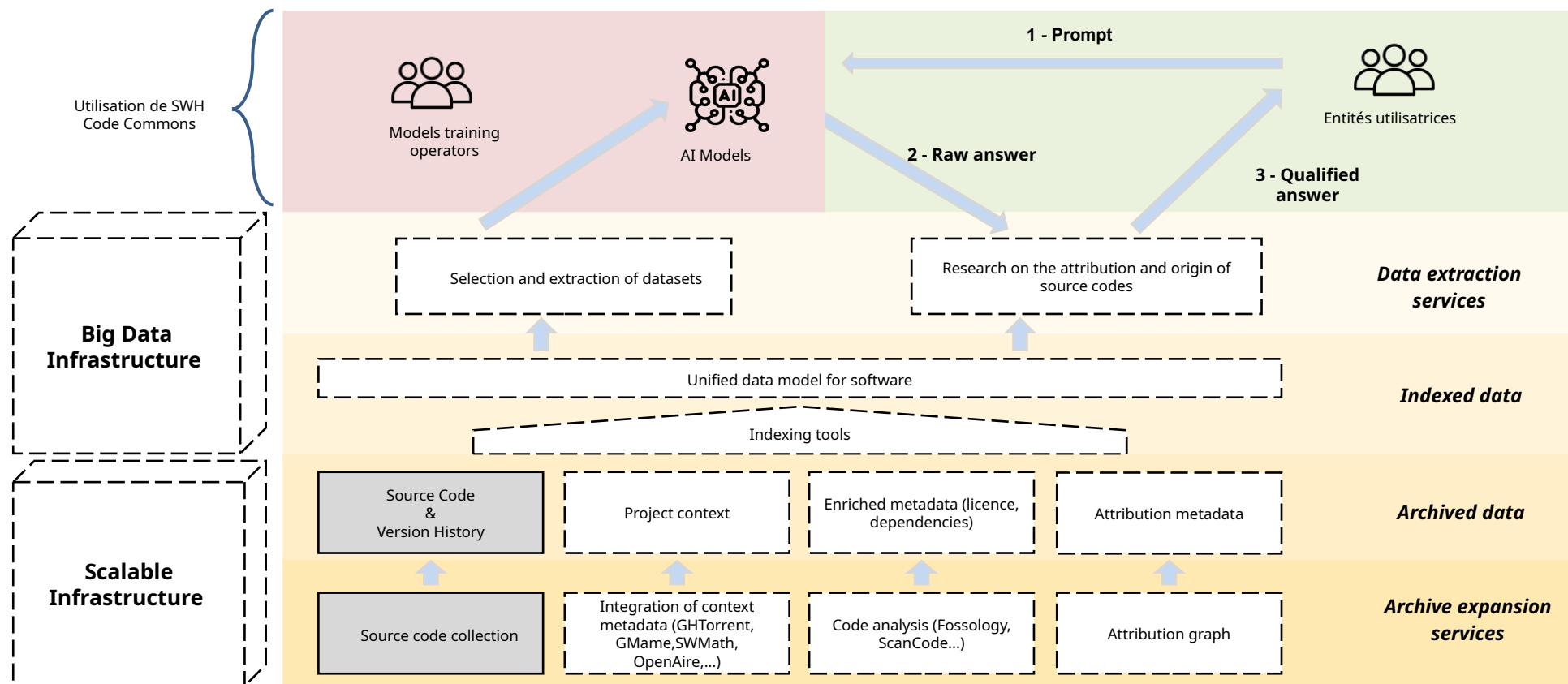
Big Data Infrastructure

Selection and extraction of datasets

Research on the attribution and origin of source codes

Data extraction services

Unified data model for software

Indexed data

Indexing tools

Scalable Infrastructure

Source Code & Version History

Project context

Enriched metadata (licence, dependencies)

Attribution metadata

Archived data

Source code collection

Integration of context metadata (GHTorrent, GMame,SWMath, OpenAire,...)

Code analysis (Fossology, ScanCode...)

Attribution graph

Archive expansion services

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

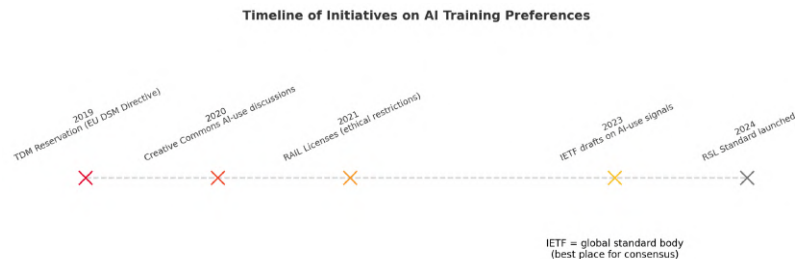# A peek at the AI Landscape: One Year Later

deepseek Smaller models with quality data may well work

Code in training data improves all models

To Code, or Not To Code? Exploring Impact of Code in Pre-training — https://arxiv.org/abs/2408.10914*At Which Training Stage Does Code Data Help LLMs Reasoning? —* https://arxiv.org/abs/2309.16298*Unveiling the Impact of Coding Data Instruction Fine-Tuning on Large Language Models Reasoning —* https://arxiv.org/abs/2405.20535

IBM **Infosys®** SCAI members and CodeCommons external contributors

Active work on AI preferences

Timeline of Initiatives on AI Training Preferences

2019 TDM Reservation (EU DSM Directive)
2020 Creative Commons AI-use discussions
2021 RAIL Licenses (ethical restrictions)
2023 IETF drafts on AI-use signals
2024 RSL Standard launched

IETF = global standard body
(best place for consensus)

AI Factory France   About   Services   News & Events   Contact

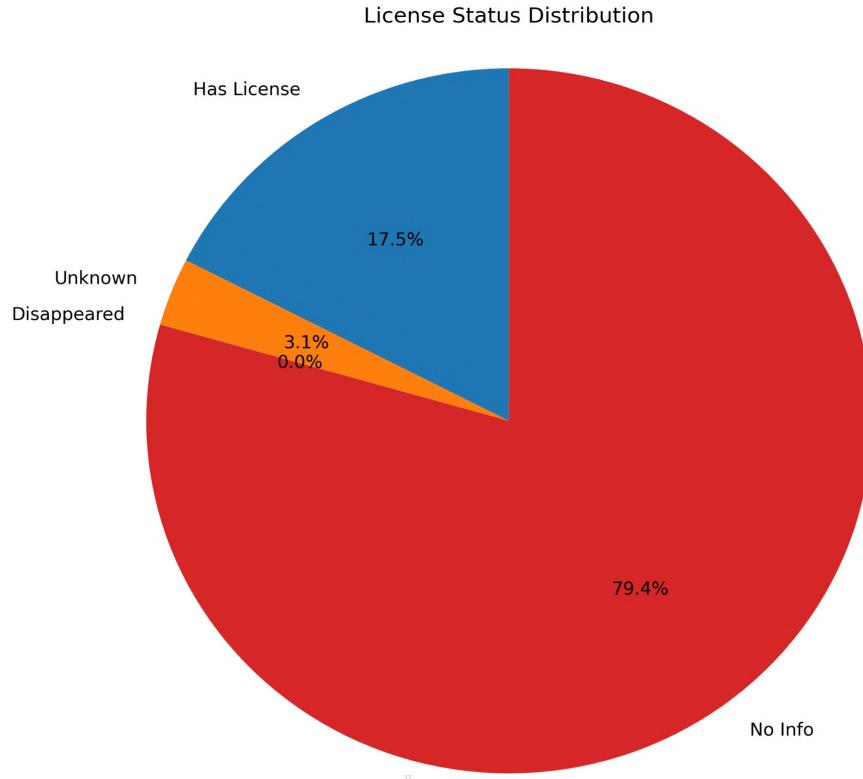**A hub to Accelerate AI for Science, Industry & Society**

AI Factory France is a national and European-scale platform backed by a broad coalition of France's most prestigious academic, public, and private partners — including GENCI, Inria, CNRS, CEA, France Universités with 12
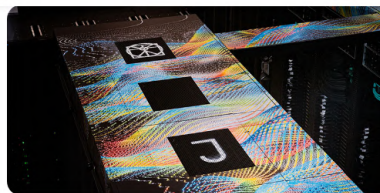
Jean Zay – Copyrights : Cyril FRESILLON / IDRIS / CNRS Images

CodeCommons is more relevant than ever!

CODE COMMONS: bird's eye view of the legal implications

What about no license?

Even with a license, what about AI preferences?

License Status Distribution

- Has License — 17.5%
- Unknown / Disappeared — 3.1% / 0.0%
- No Info — 79.4%

Creative Commons Signals:
https://creativecommons.org/2025/06/25/introducing-cc-signals-a-new-social-contract-forthe-age-of-ai/

IETF aipref WG:
https://datatracker.ietf.org/wg/aipref/about/

RSL Standard: https://rslstandard.org/
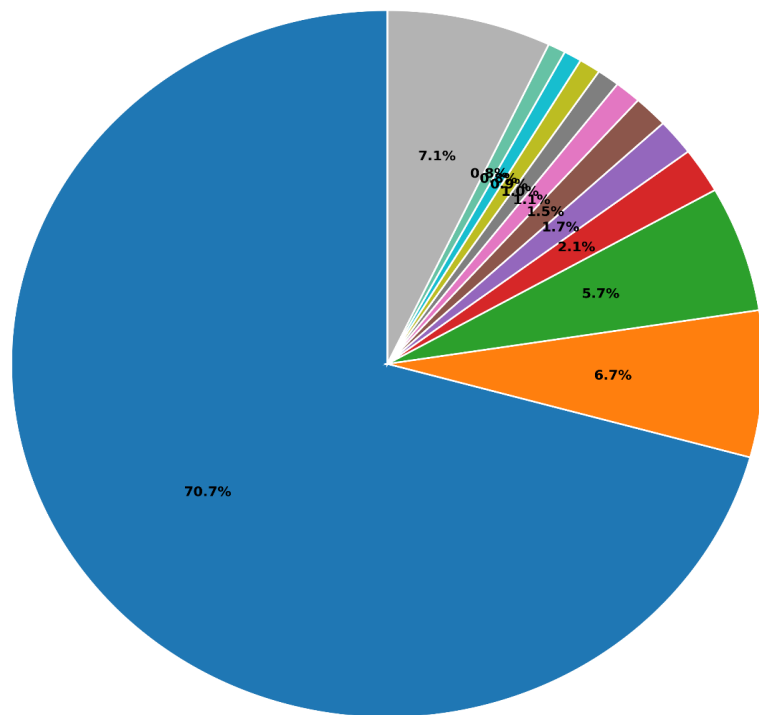
They focus on Web, we miss source code!

Need to engage now →

A look at Open Source infrastructure dependencies

What development platforms in France?

Distribution of Project Platforms (Base URLs)

Pie chart values: 70.7%, 6.7%, 5.7%, 2.1%, 1.7%, 1.5%, 1.1%, 0.8%, 7.1%

What happens if we lose access?

We need focused, massive engagement to create a key infrastructure at European level, on top of Software Heritage for:

- R E S I L I E N C E
- C Y B E R S E C U R I T Y
- T R A N S P A R E N C Y I N AI
- L A R G E S C A L E SOFTWARE STUDIES

The time is now!