

From Software Heritage to Code Commons

a vision for transparent and responsible AI
in code-based model training

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



Closed model 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 model 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

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

Let's look at the case of software source code



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

availability

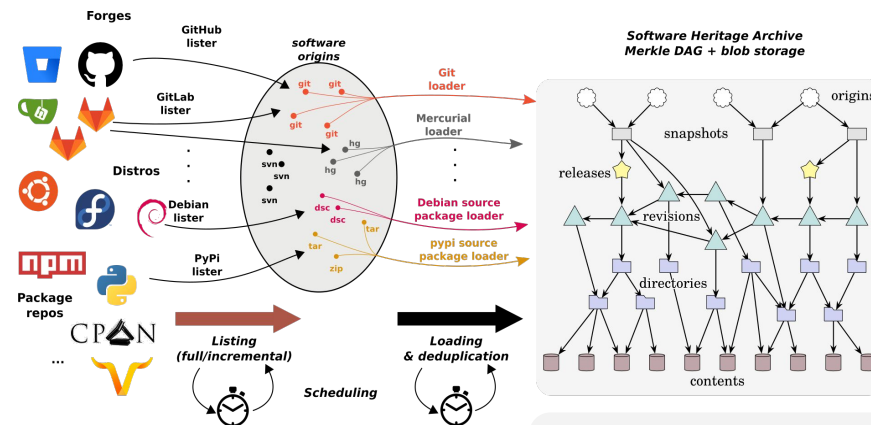
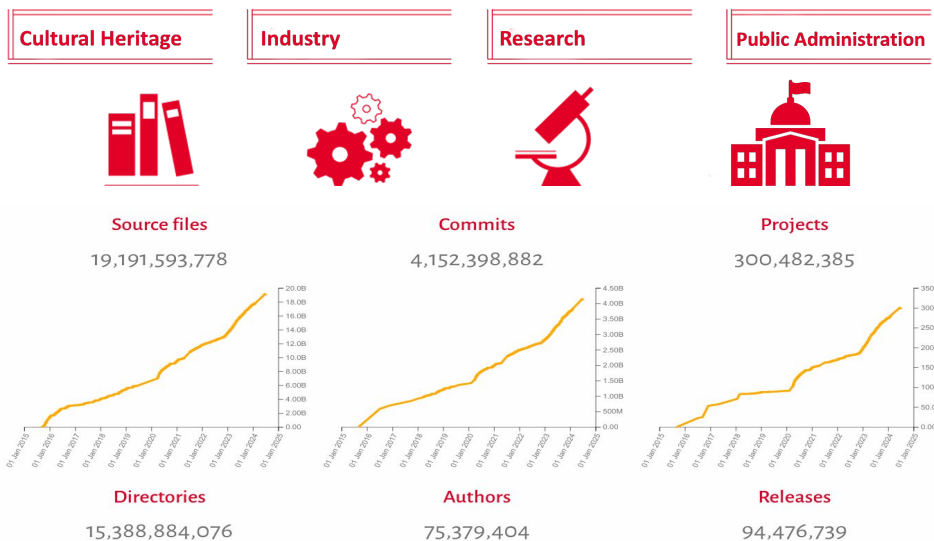
transparency

integrity

Starting point

The largest open source code archive

Unique digital common good *built in France since 2015*



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

A **unique infrastructure** of its kind
to meet the needs of AI...



Integrity and identification of 50B+ artifacts in the archive

Software Hash Identifiers (ISO 18670)

ISO Standards Sectors About ISO Insights & news

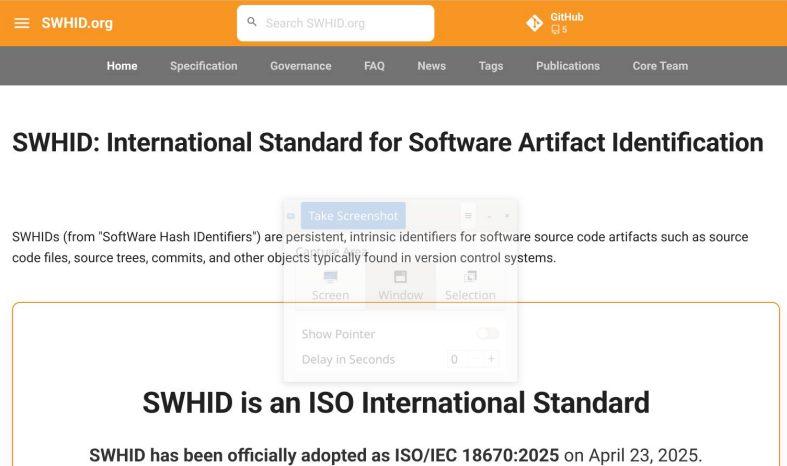


ISO/IEC 18670:2025
Information technology —
SoftWare Hash Identifier
(SWHID) Specification V1.2

Read sample

Published (Edition 1, 2025)

white paper on identifiers for the CRA:
<https://hal.science/hal-05009757>



SWHID.org

Home Specification Governance FAQ News Tags Publications Core Team

SWHID: International Standard for Software Artifact Identification

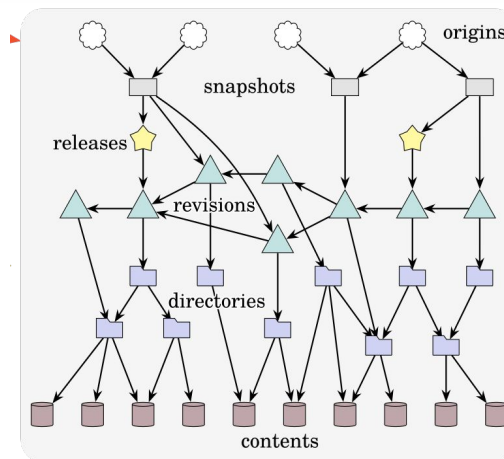
SWHIDs (from "SoftWare Hash Identifiers") are persistent, intrinsic identifiers for software source code artifacts such as source code files, source trees, commits, and other objects typically found in version control systems.

SWHID is an ISO International Standard

SWHID has been officially adopted as ISO/IEC 18670:2025 on April 23, 2025.

AI anecdote

Traceability

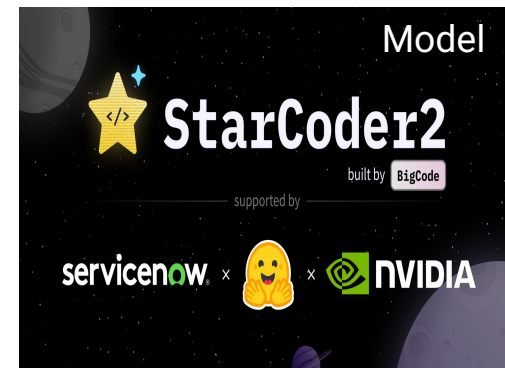
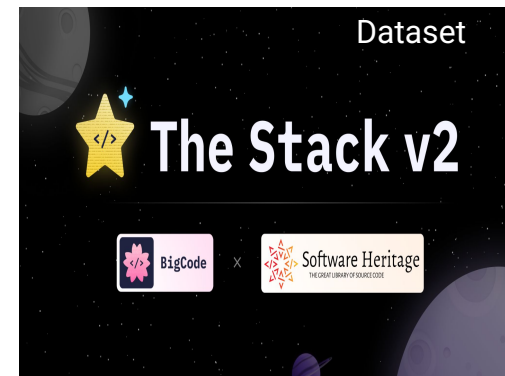


Looking for founding principles at Software Heritage

Findings from BigCode: The Stack v2 and StarCoder2

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 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.



Yes, it's possible

but it's hard to do it well

© October 19, 2023

Software Heritage Statement on Large Language Models for Code



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 fairly 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 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

today!

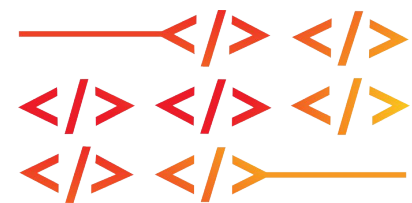


Need a **coordinated effort** to address these issues

It's time to build a

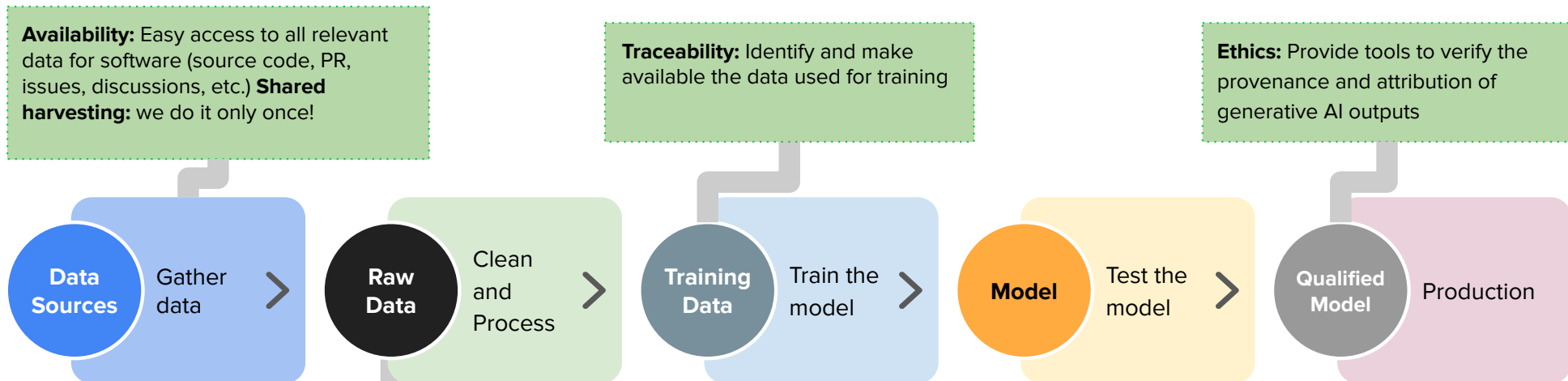
Code Commons





CODE COMMONS

5Me FR funding
30 Months



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

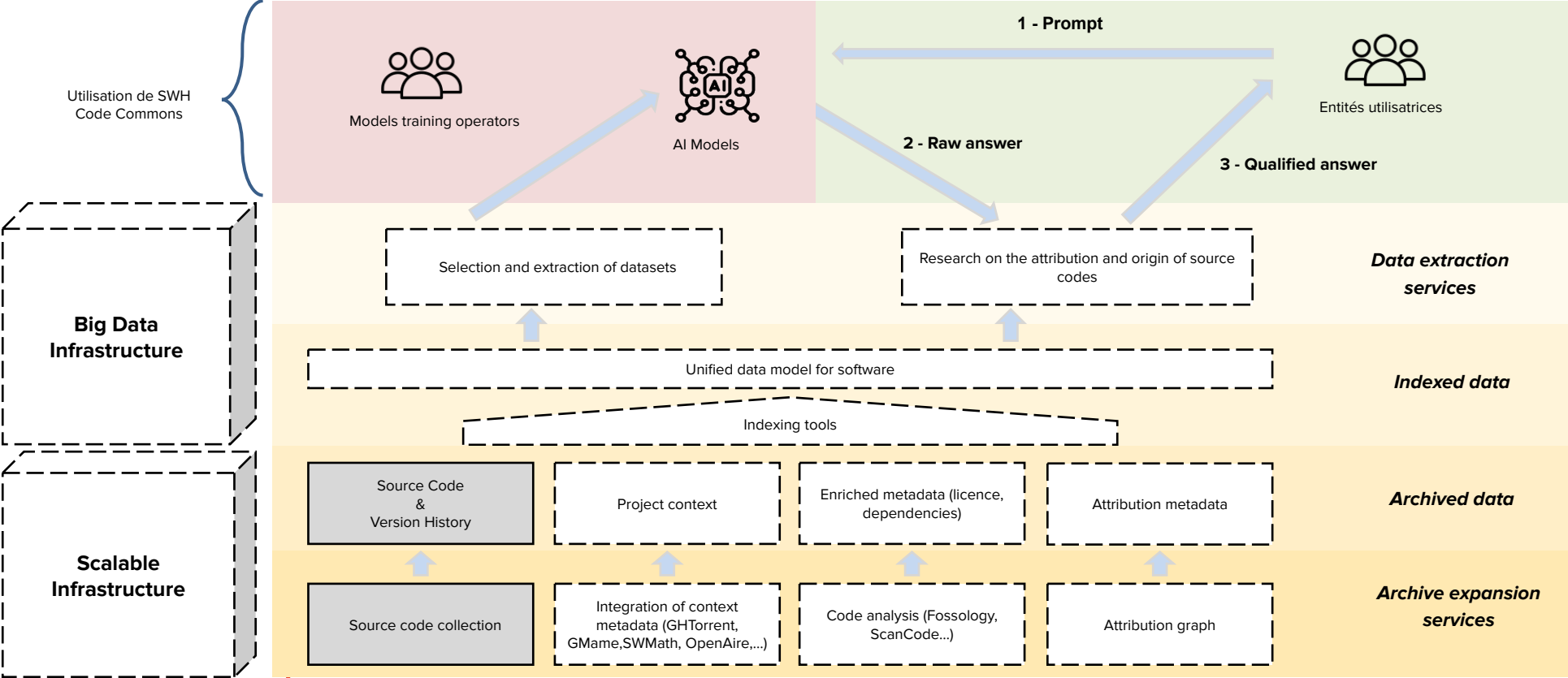
Structuring: Organize and connect the various data sources to create a coherent training set.

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

Solutions

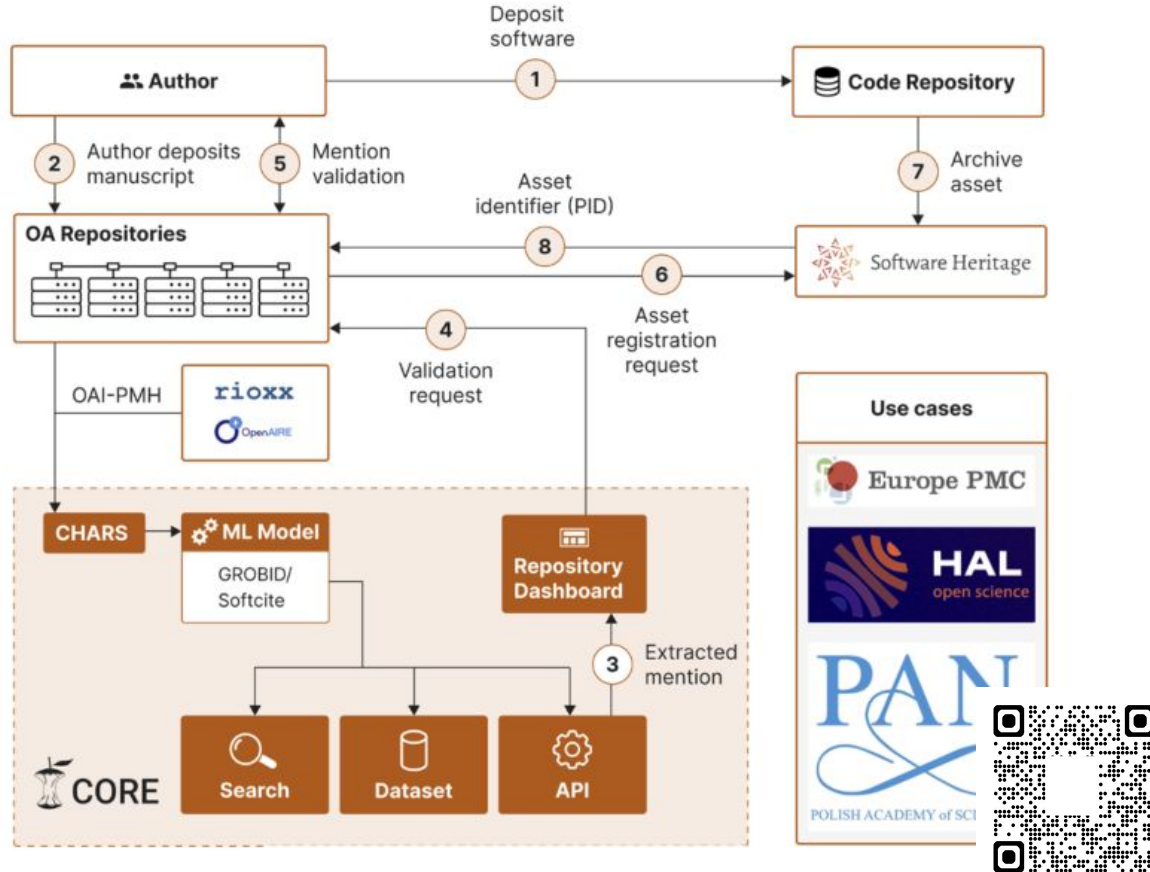


CODE COMMONS : BIRD'S EYE VIEW (technology+data spaces)



Related projects feeding the archive expansion

SoFAIR



SWH-Security



A peek at the AI Landscape: One Year Later



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>

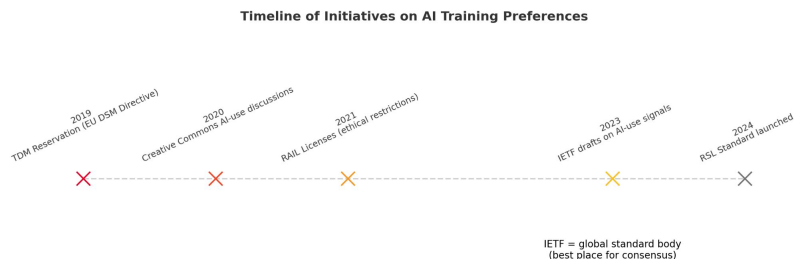
At Which Training Stage Does Code Data Help LLMs Reasoning? — <https://arxiv.org/abs/2309.16298>

Unveiling the Impact of Coding Data Instruction Fine-Tuning on Large Language Models Reasoning — <https://arxiv.org/abs/2405.20535>



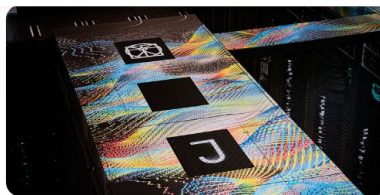
SCAI members and CodeCommons external contributors

Active work on AI preferences



AI Factory France

[About](#) [Services](#) [News & Events](#) [Contact](#)



Jean Zap - Copyright - GYM FRESILLON / IDRS / CNRS Images

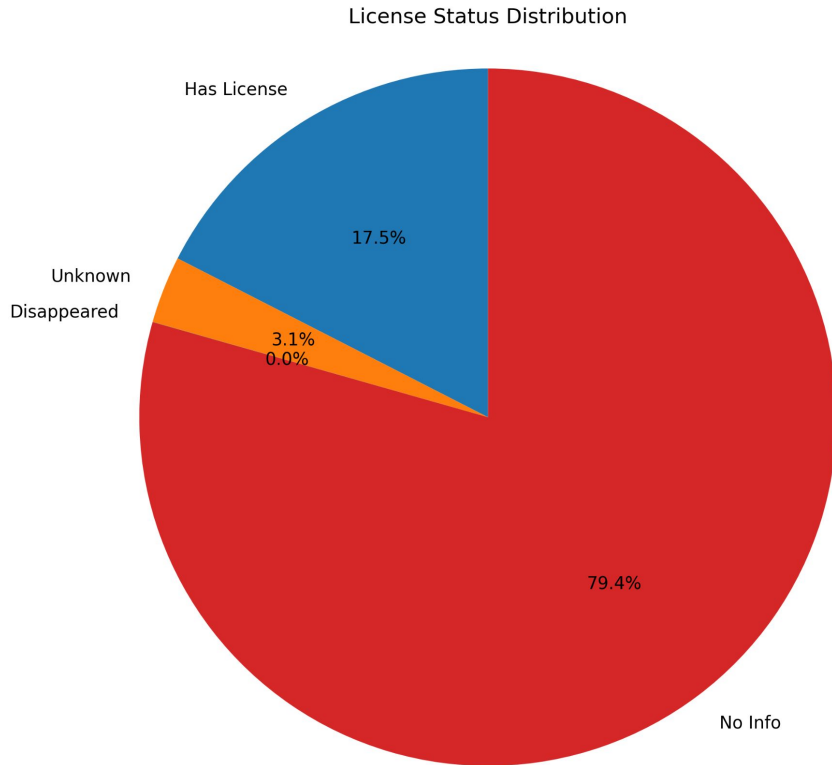
A hub to Accelerate AI for Science, Industry & Society

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

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?

Creative Commons Signals:

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

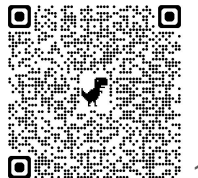
IETF aipref WG:

<https://datatracker.ietf.org/wg/aipref/about/>

RSL Standard: <https://rslstandard.org/>

They focus on Web, we miss source code!

Need to engage now →



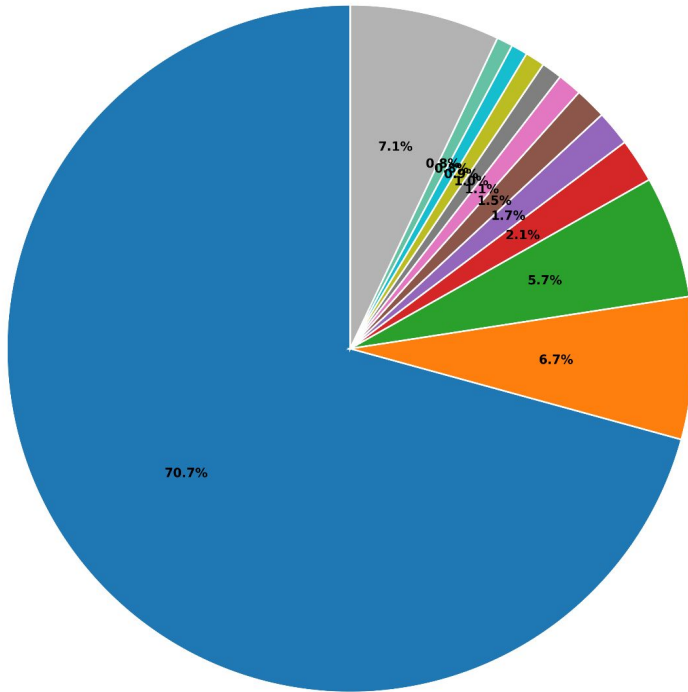
1

A look at Open Source infrastructure dependencies

What development platforms in France?

What happens if we lose access?

Distribution of Project Platforms (Base URLs)



We need focused, massive engagement to create a

RESILIENCE infrastructure

at the European level

The time is now!