No Knowledge without Source

Collecting, Preserving and Sharing Software in a Risky World

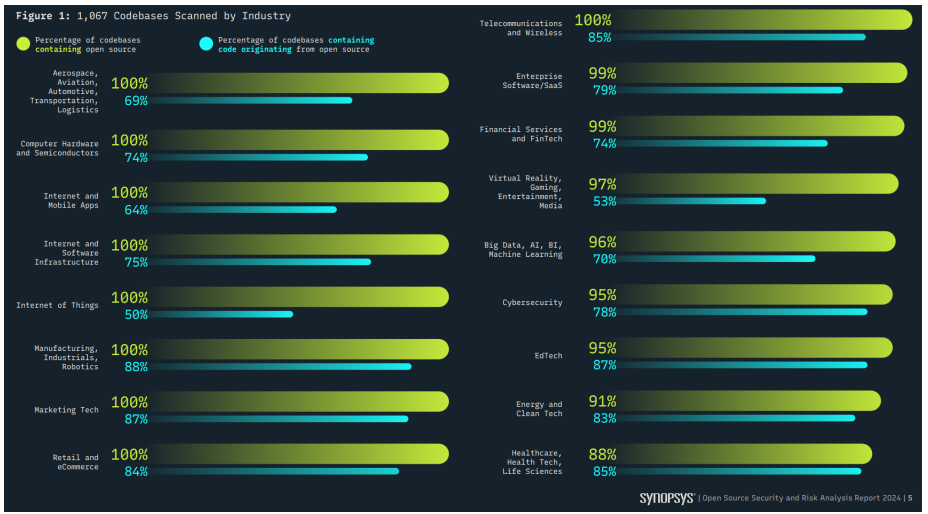
Roberto Di Cosmo

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

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

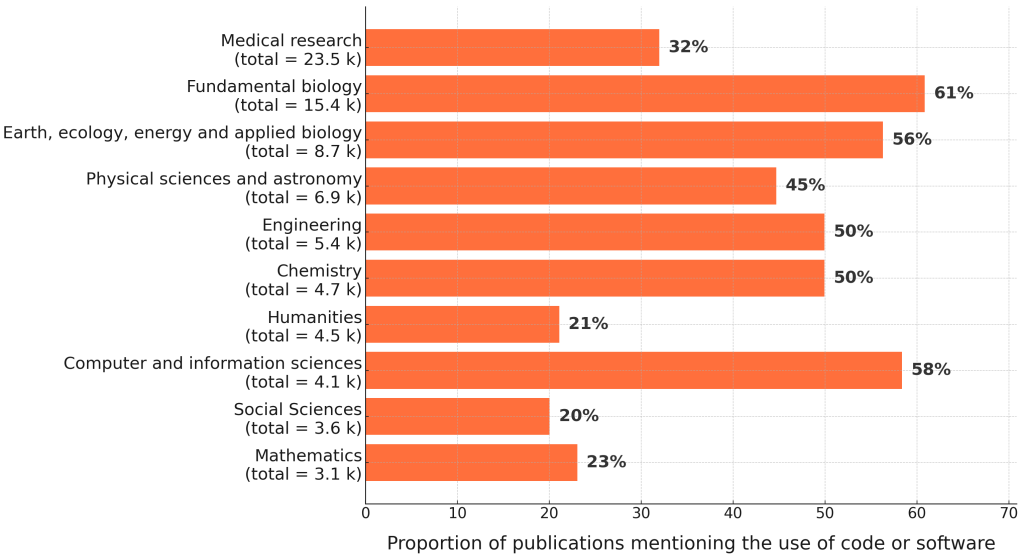
Industry



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

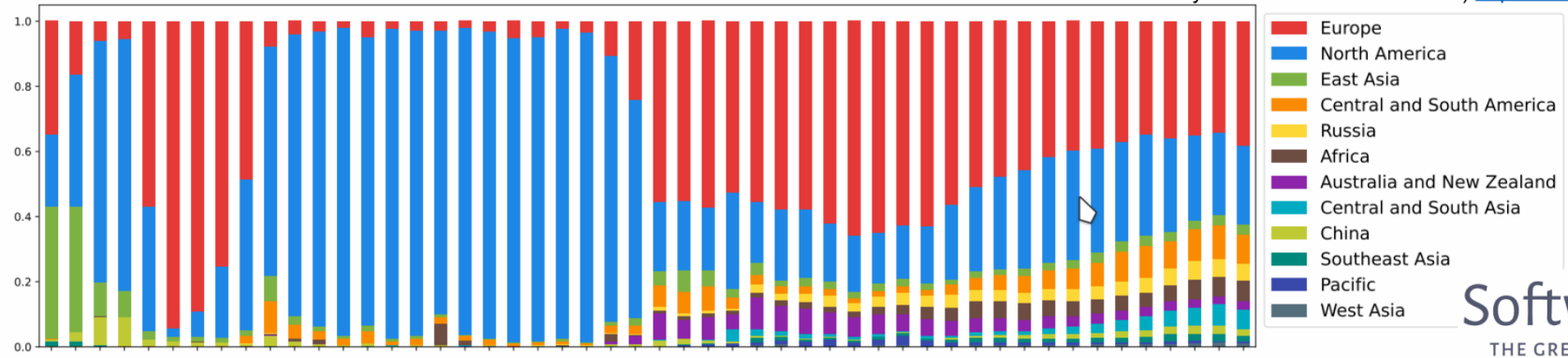


French Open Science Monitor 2025

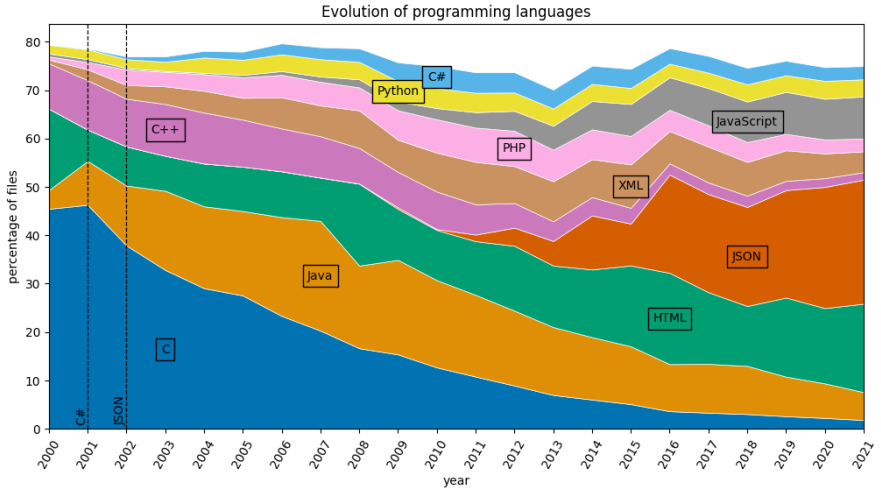
Open Source Software: global and growing

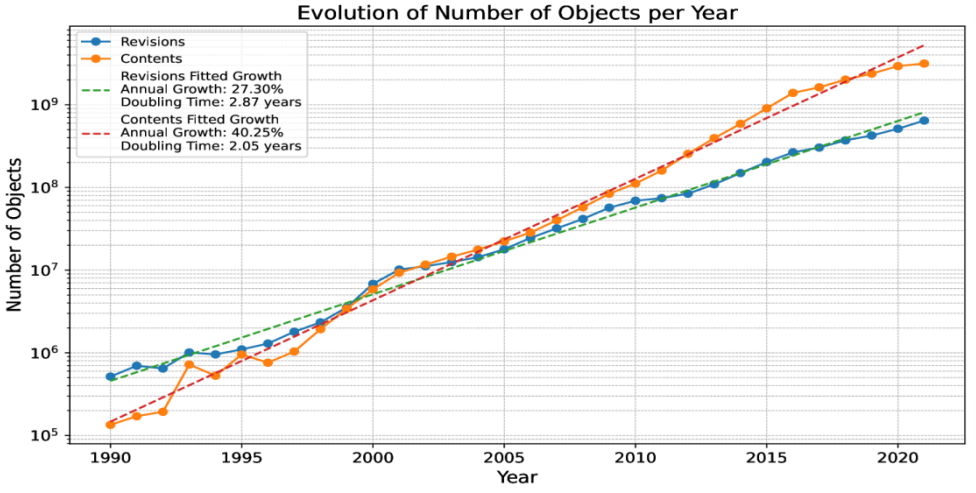
Ratio of commits by world zone over the 1971–2020 period.

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



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE





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

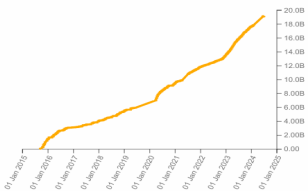
Inria with unesco

The largest open source code archive

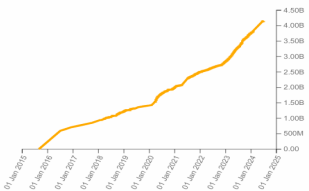
Unique digital common good *built in France since 2015*

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

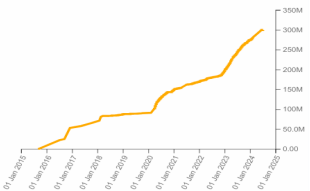
Source files
19,191,593,778



Commits
4,152,398,882



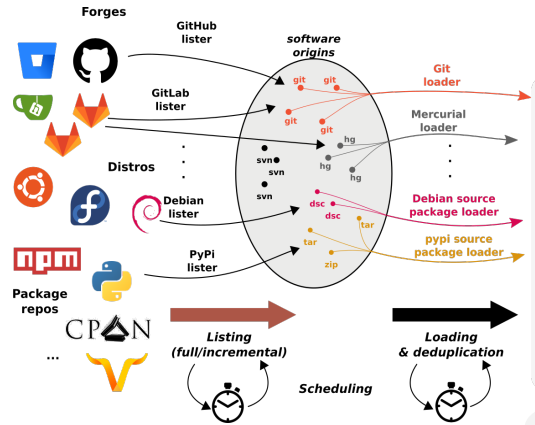
Projects
300,482,385



Directories
15,388,884,076

Authors
75,379,404

Releases
94,476,739



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**

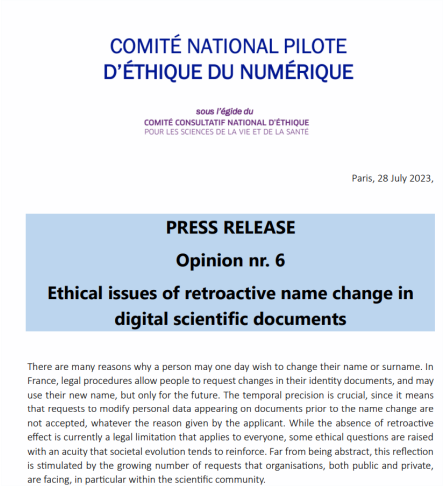
} of source codes

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

More risks and challenges

Collateral damage from

- EU Copyright Directive
- GDPR



COMITÉ NATIONAL PILOTE
D'ÉTHIQUE DU NUMÉRIQUE

sous l'égide du
COMITÉ CONSULTATIF NATIONAL D'ÉTHIQUE
POUR LES SCIENCES DE LA VIE ET DE LA SANTE

Paris, 28 July 2023,

PRESS RELEASE

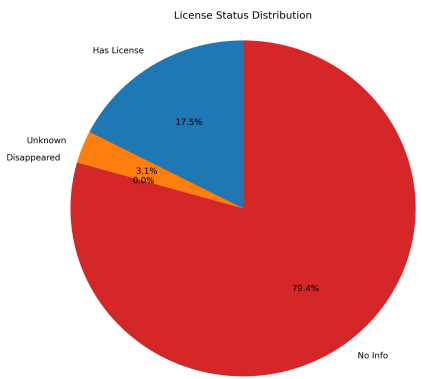
Opinion nr. 6

Ethical issues of retroactive name change in digital scientific documents

There are many reasons why a person may one day wish to change their name or surname. In France, legal procedures allow people to request changes in their identity documents, and may use their new name, but only for the future. The temporal precision is crucial, since it means that requests to modify personal data appearing on documents prior to the name change are not accepted, whatever the reason given by the applicant. While the absence of retroactive effect is currently a legal limitation that applies to everyone, some ethical questions are raised with an acuity that societal evolution tends to reinforce. Far from being abstract, this reflection is stimulated by the growing number of requests that organisations, both public and private, are facing, in particular within the scientific community.

New obligations

e.g. Cyber resilience act (ongoing, you should look at it)



Legal uncertainty: vast majority of publicly available code is **shared** to be reused, but has **no licence** information

Coping with the (Gen)AI tidal wave

Looking for founding principles at Software Heritage

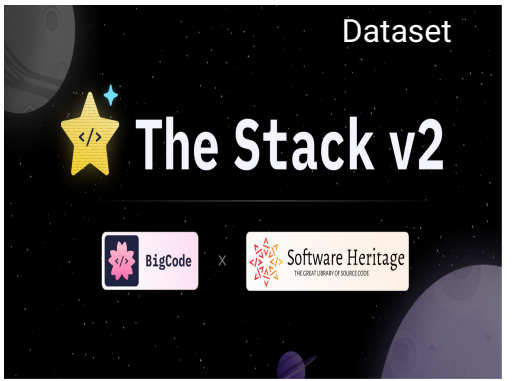
Findings from BigCode:
The Stack v2 and StarCoder2

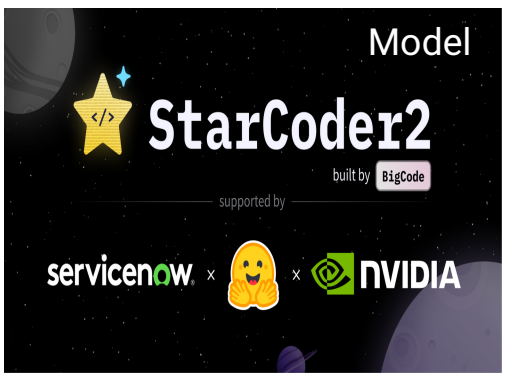
October 19, 2023

Software Heritage Statement on Large Language Models for Code

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.

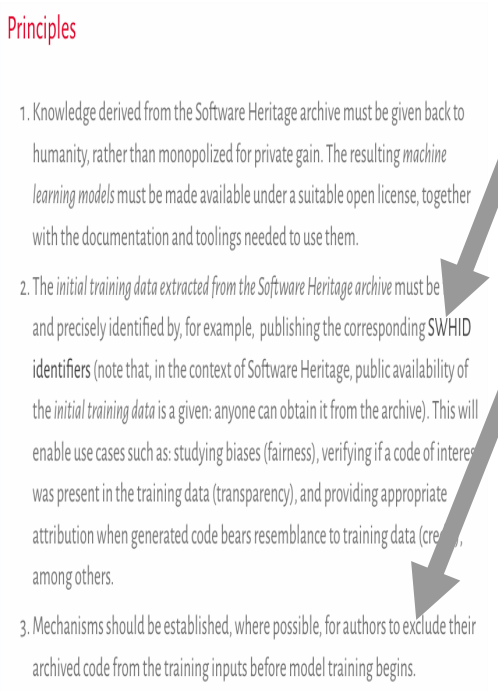




Yes, it's possible

but it's hard to do it well

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 interes was present in the training data (transparency), and providing appropriate attribution when generated code bears resemblance to training data (cre , 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

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

Software Heritage Membership Program Source Code & AI

today!

(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!)

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

It's time to build a

Code Commons

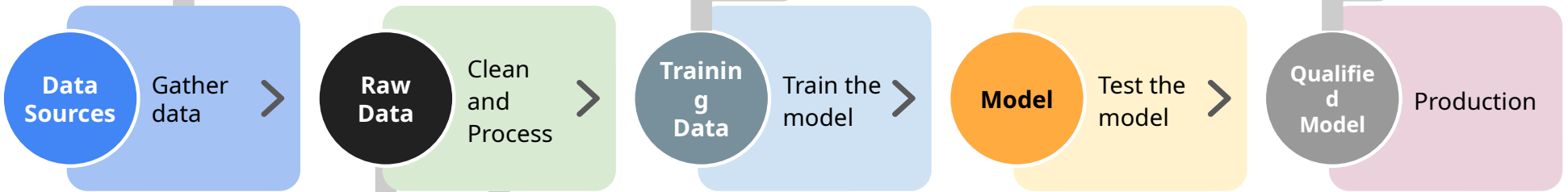
CODE COMMONS

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



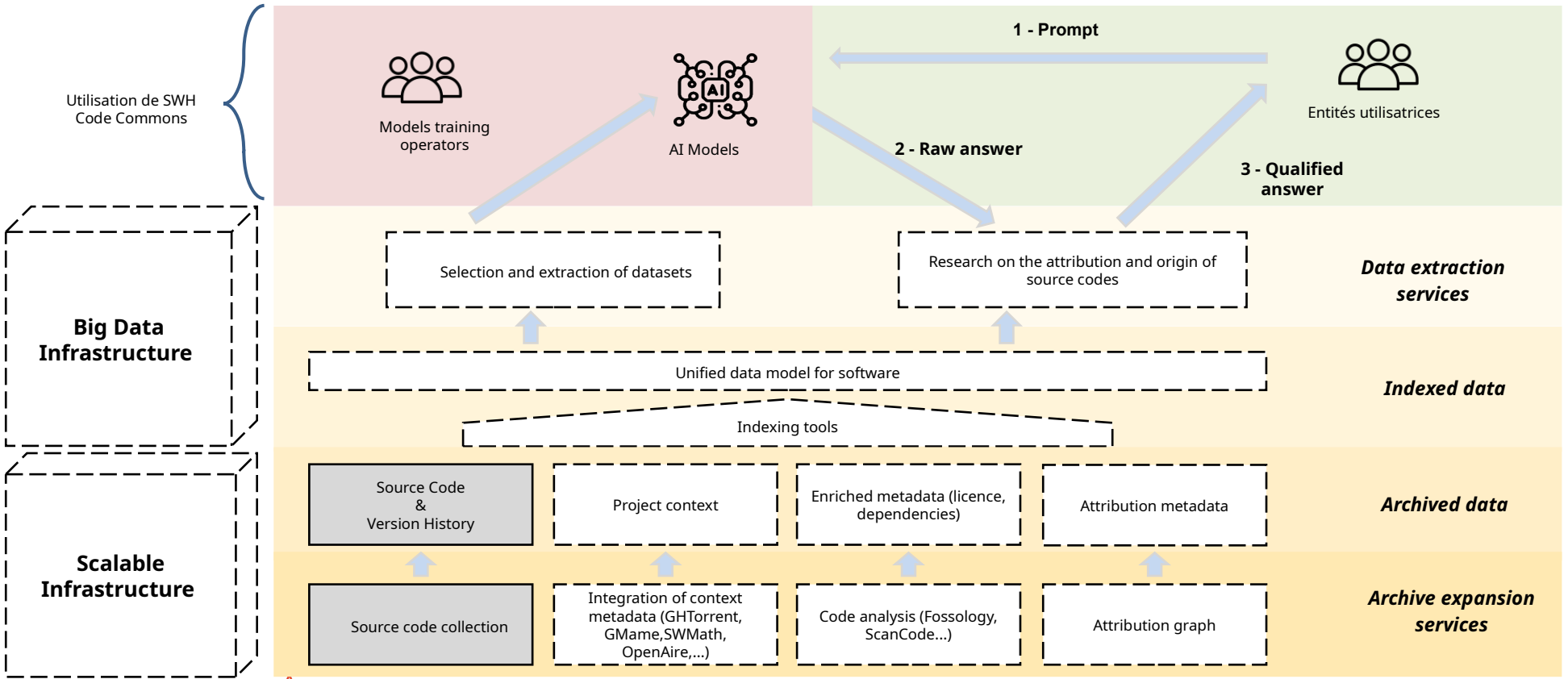
Structuring: Organize and connect the various data sources to create a coherent training set.

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

Solutions

Inria · Software Heritage · CEA · TWEAG by Modus Create · Sant'Anna School of Advanced Studies – Pisa · Università degli Studi di Torino · Università di Pisa · Alma Mater Studiorum Università di Bologna · AboutCode · GENCI HPC at the service of knowledge · bpifrance SERVIR L'AVENIR · FRANCE 2030

CODE COMMONS : BIRD'S EYE VIEW (technology)



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

A peek at the AI Landscape: One Year Later

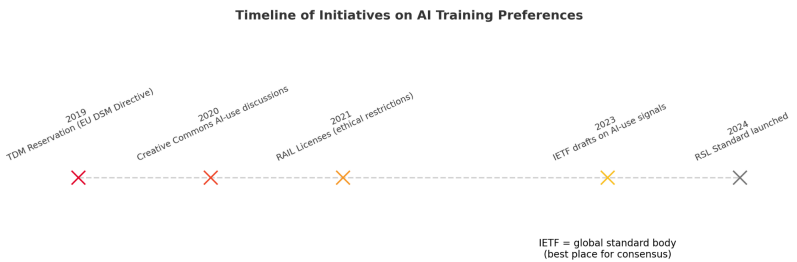
deepseek Smaller models with quality data may well work

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

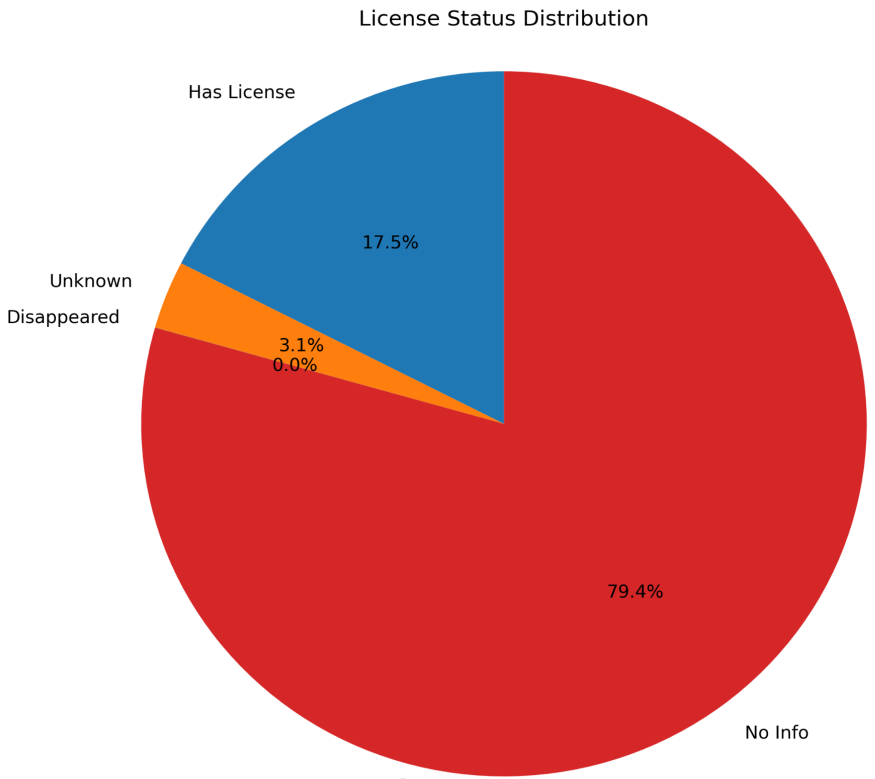
Active work on AI preferences



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/

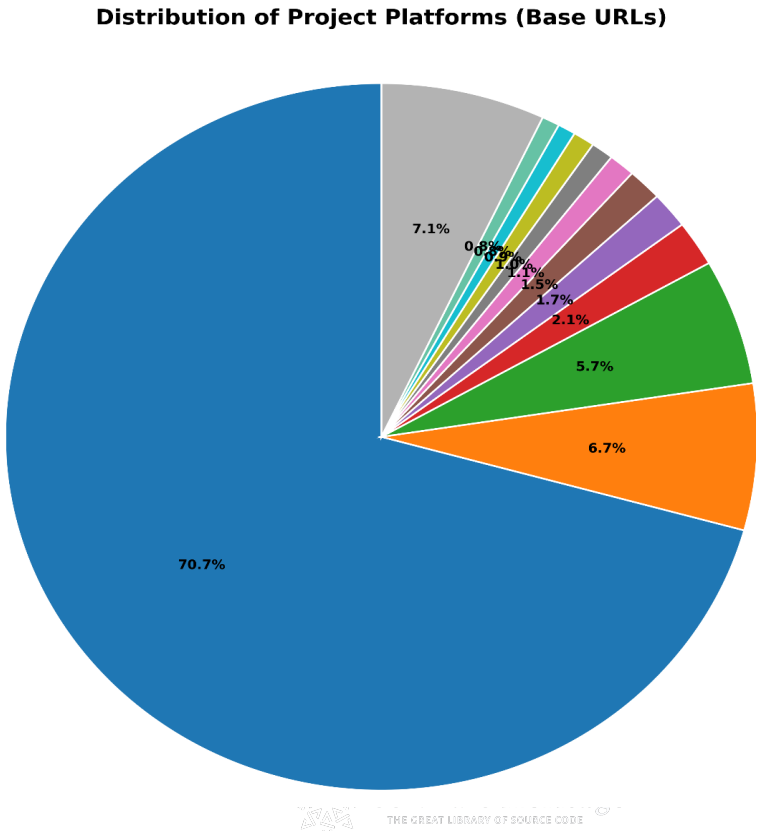
RSL Standard: https://rslstandard.org/

They focus on Web, we miss source code!

Need to engage now →

A look at Open Source infrastructure dependencies

What development platforms in France?



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:

- **RESILIENCE**
- **CYBERSECURITY**
- **TRANSPARENCY IN AI**
- **LARGE SCALE SOFTWARE STUDIES**

The time is now!

Appendix

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

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

availability

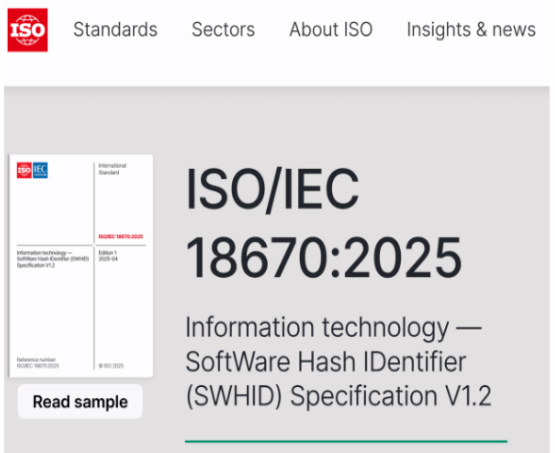
transparency

integrity

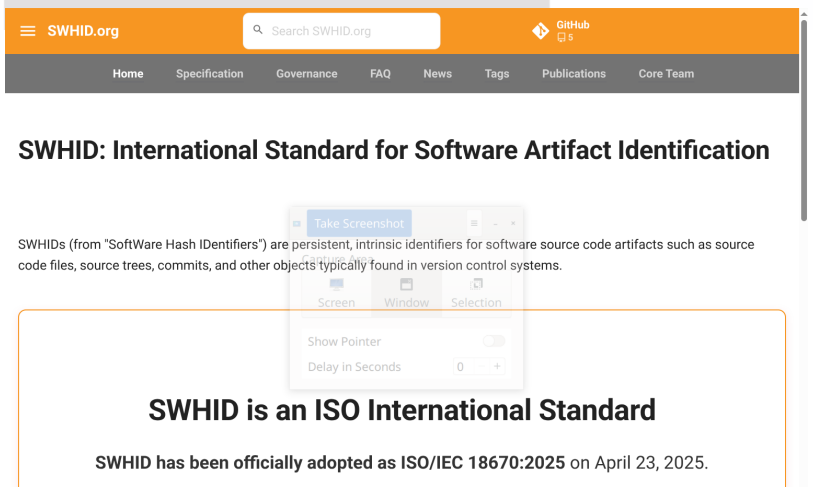
Let's look at the case of software source code

Integrity and identification of 50B+ artifacts in the archive

Software Hash Identifiers (ISO 18670)

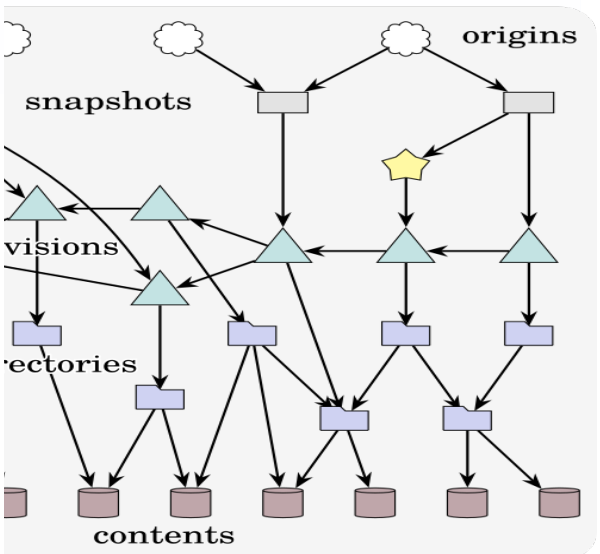


white paper on identifiers for the CRA:
https://hal.science/hal-05009757



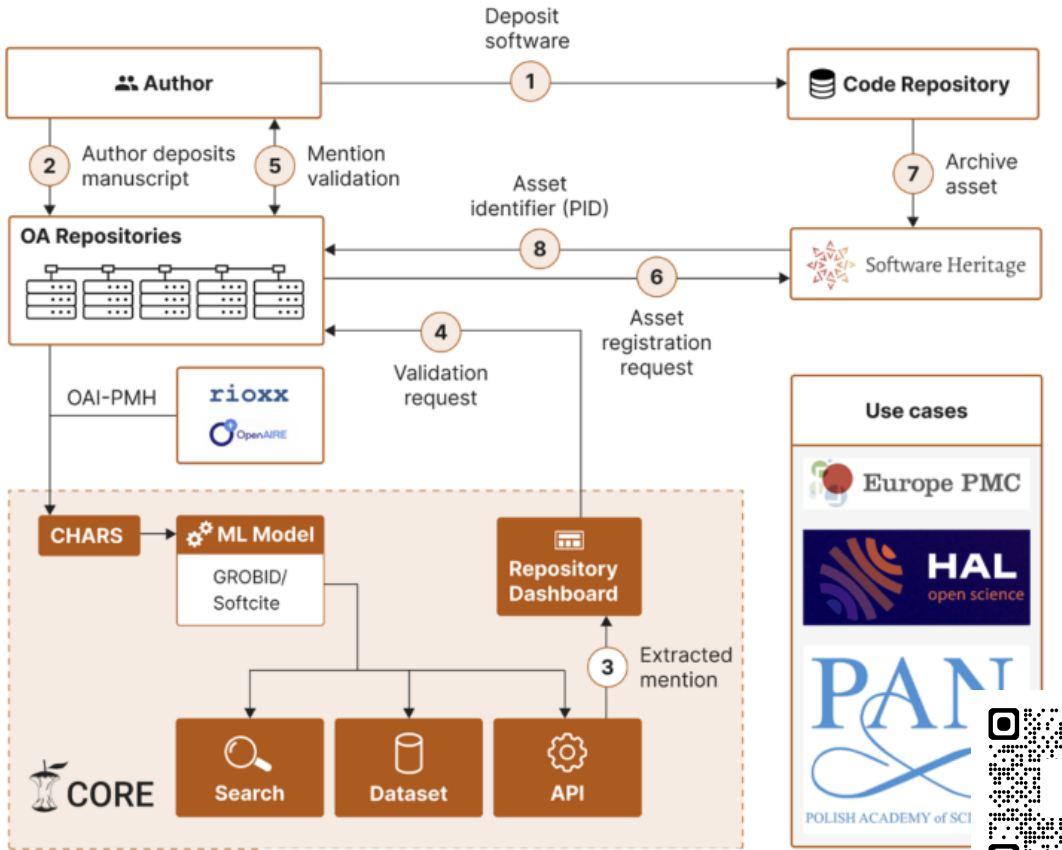
AI anecdote

Traceability



Related projects feeding the archive expansion

SoFAIR



SWH-Security