From **Software Heritage** to **CodeCommons**:

a vision for *transparent and responsible* ***data*** for AI

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

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

availability

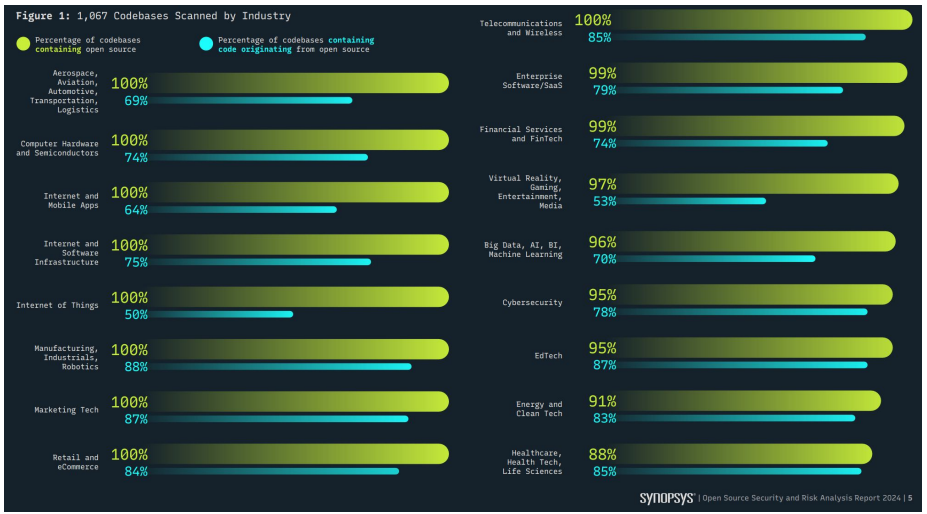
transparency

integrity

Let's look at the case of software source code

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

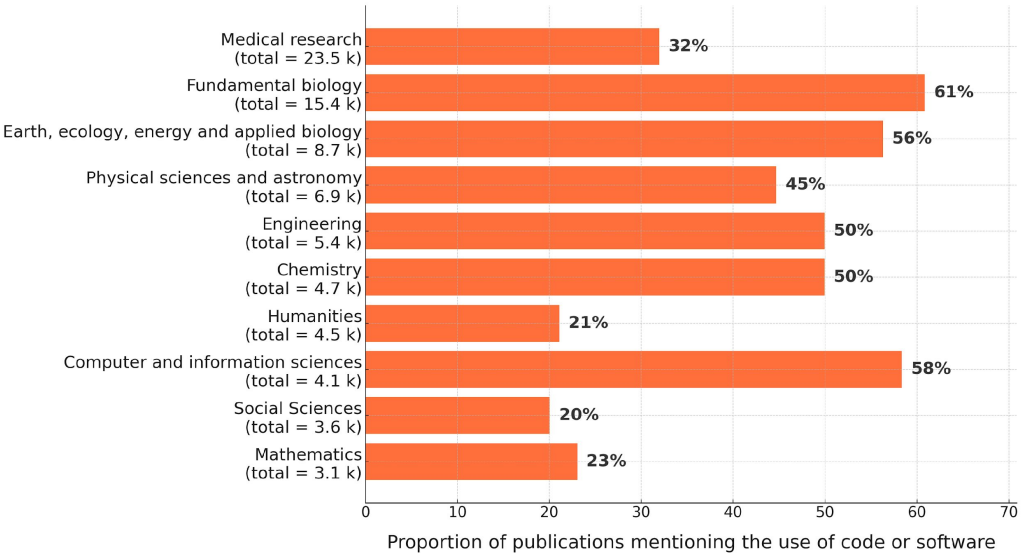
Industry



- **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

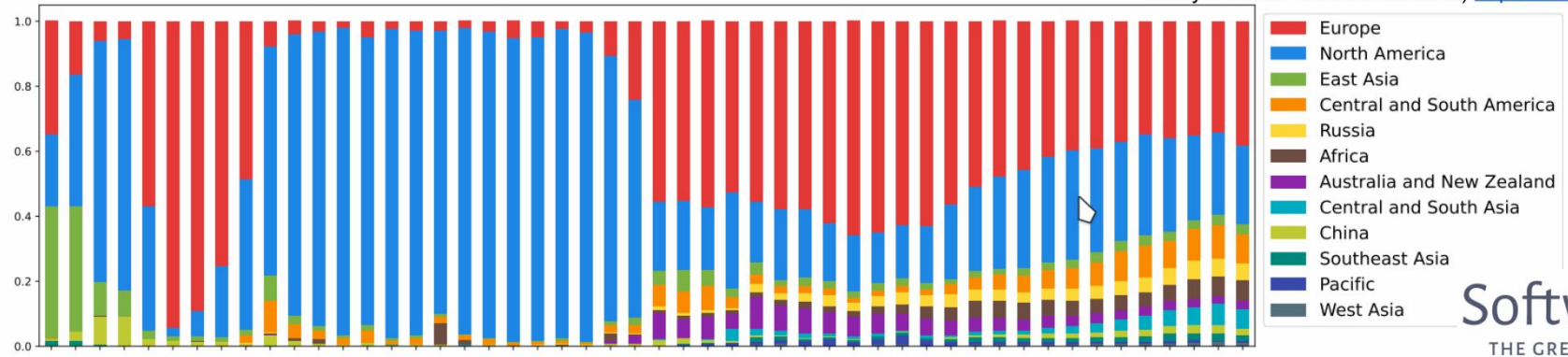


French Open Science Monitor 2025

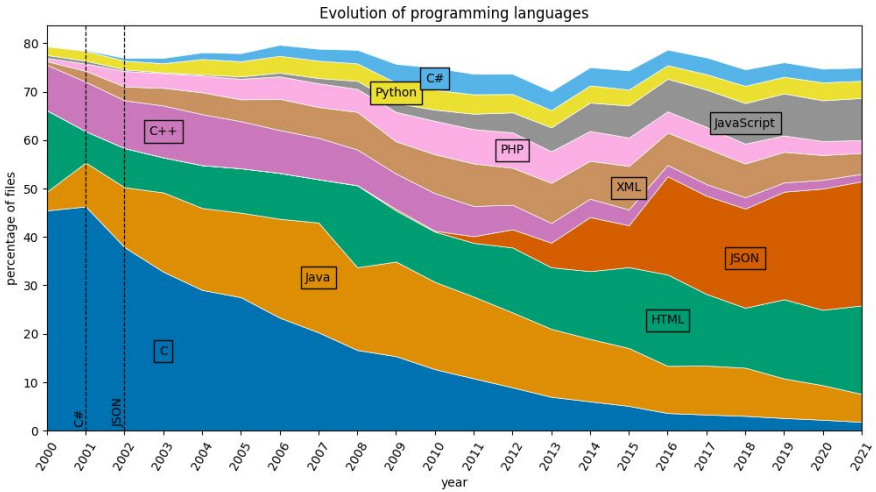
Publicly available Software Source Code: global and growing

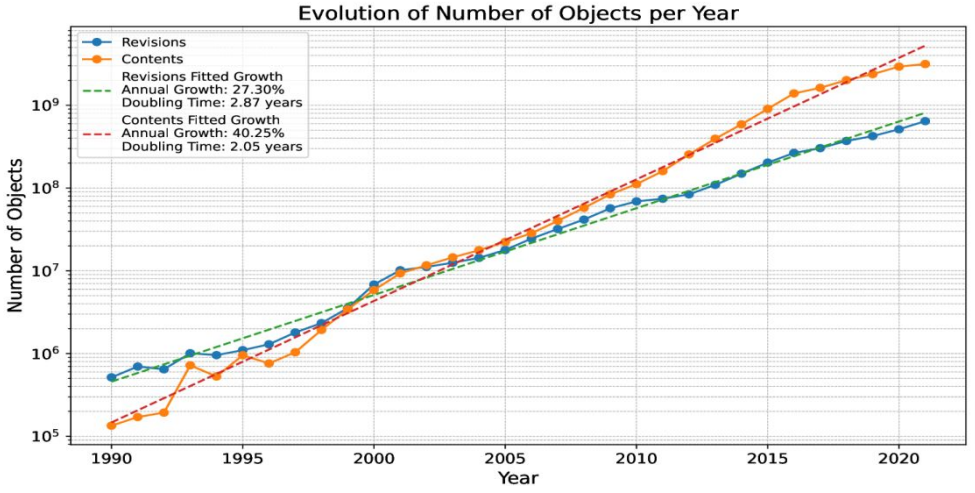
Ratio of commits by world zone over the 1971–2020 period.

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.

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Inria with unesco

Software Heritage
A decade of achievement
Activity Report 2025

The largest open source code archive

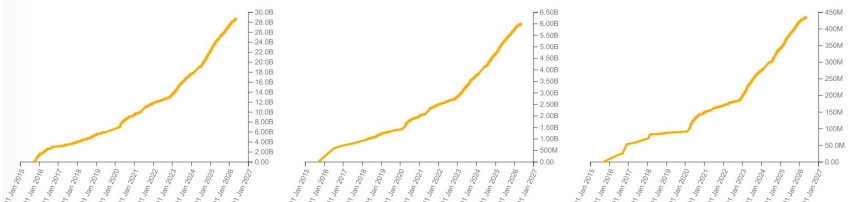
Unique digital common good *built in France* ***since 2015***

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

Source files
28,782,784,702

Commits
6,017,300,328

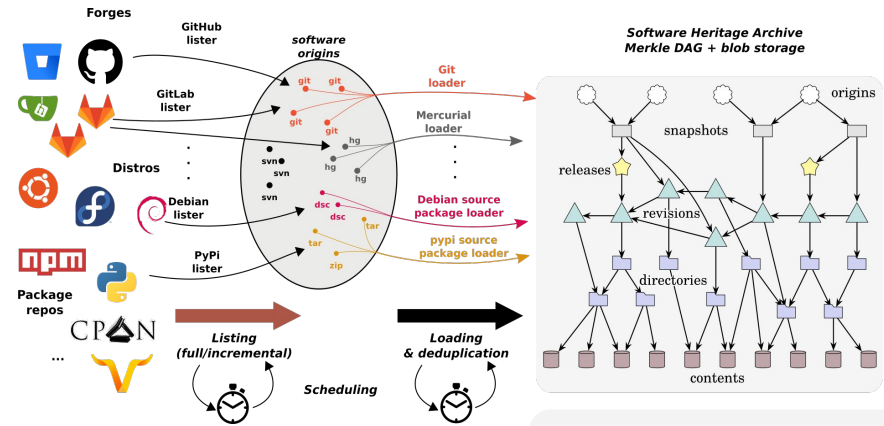
Projects
435,822,663



Directories
22,433,217,096

Authors
106,918,400

Releases
161,037,805



5000+ platforms

All versions, all history development in a single graph

- **50 × 10⁹** nodes
- **1000 × 10⁹** edges
- ~ **6 PB** of storage

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

A **unique infrastructure** to support
transparent and responsible
source code **sourcing** for Big Data and **AI**

Coping with the (Gen)AI tidal wave

Looking for founding principles at Software Heritage

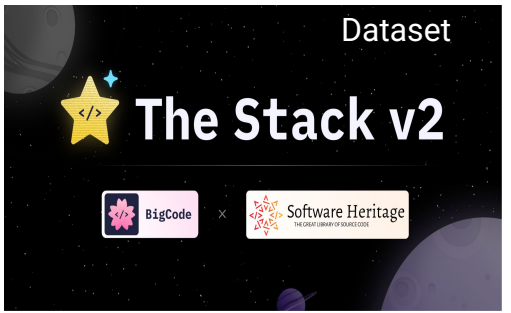
Findings from BigCode: The Stack v2 and StarCoder2

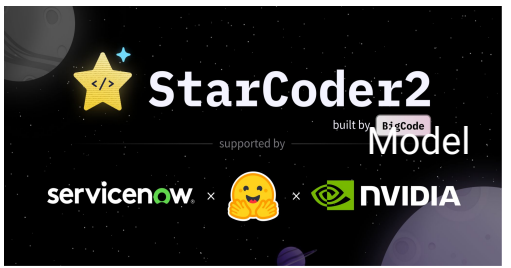
October 19, 2023

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.





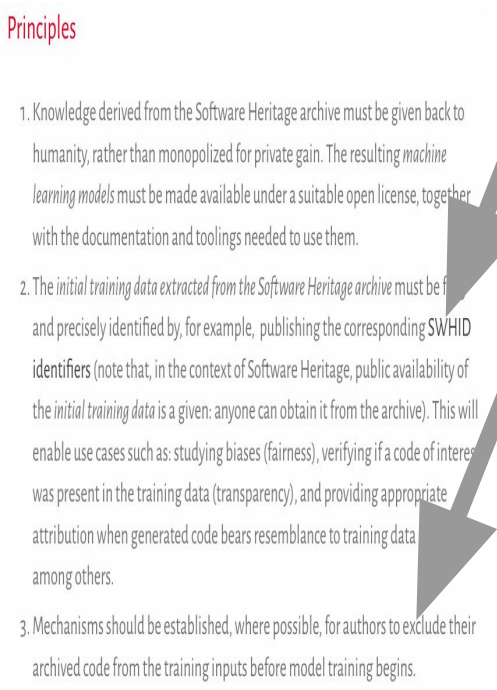
You are all using the Stack V2

Who is:

Respecting the principles?

Contributing back?

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 f 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 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 (ISO/IEC 18670)
and Software Heritage: ***no need to share petabytes of data***!

(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!

Software Heritage Membership Program Source Code & AI

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

It's time to build a

Code Commons

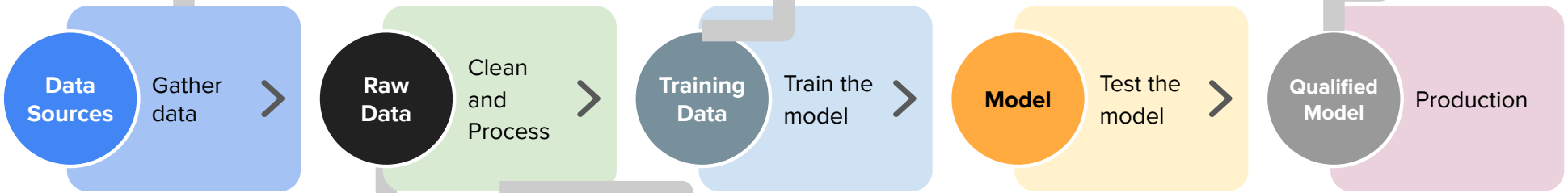
CODE COMMONS

5Me FR funding
32 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

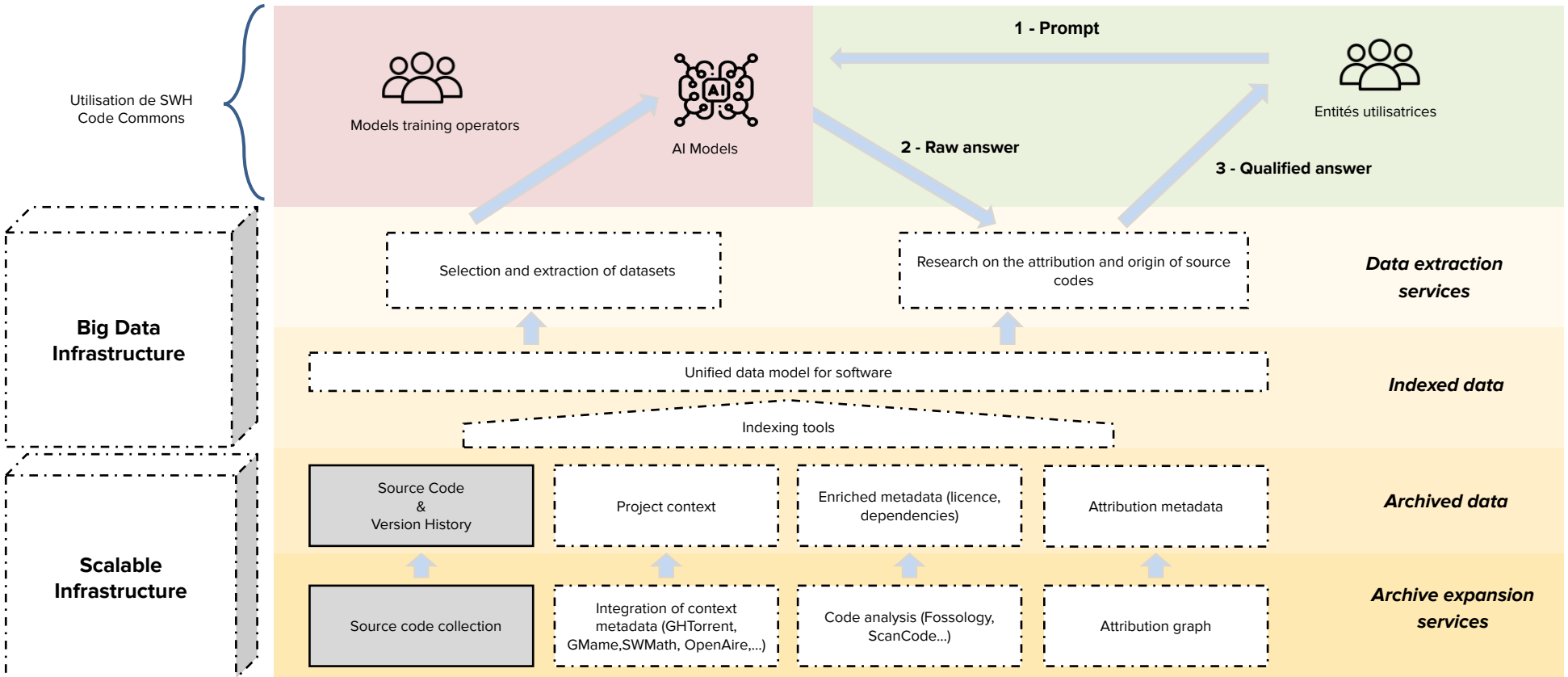


Structuring: Organize and connect the various data sources to create a coherent training set.

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

Solutions

CODE COMMONS : BIRD'S EYE VIEW (technology)



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

SCAI members today

IBM Infosys TII Technology Innovation Institute

AI preferences proposal: automatic opt-out support for source code

First prototypes for semantic similarity code search and exploration.

CodeCommons makes Software Heritage Your* unique data broker *for* responsible, efficient, transparent and accountable *sourcing of source code.

Software Heritage is key infrastructure for AI, Open Science, CyberSecurity, Resilience …

En matière de climat, de connaissances de la planète, en matière de sciences du vivant, nous avons besoin de poursuivre ce travail, de consolider les plateformes et d'avancer. Réfléchissons collectivement à la lourde tâche d'archivage de certains savoirs spécifiques. Je pense aux travaux aussi de *Software Heritage*, qui, en France, archive l'ensemble des logiciels créés dans le monde, avec une initiative européenne à construire sur le sujet. Je souhaite, au-delà de la question de notre capacité, à attirer chercheurs, enseignants-chercheurs, que nous ayons aussi une capacité à préserver, consolider, attirer plateformes de collaboration et données au service de l'intérêt général.

May 5th 2025 : Choose Europe for Science – Resilience infrastructure

Worldwide outreach

Open Science and AI
CEPAL, Chile

High level meeting at CEPAL in Chile

Open Science
Cyber and AI
Delhi, India

Open Science,
Compliance, AI
Tokyo, Japan

OSPO for Good
United Nations
New York, USA

Software Heritage + Code Commons: foundation layer for source code

Tapestry's sovereignty code-data layer must be **transparent, available**, and under **neutral governance**

It already exists!

The foundation is already here.

- **Software Heritage:** 50B+ source-code artefacts, 1 trillion edges, 5,000+ forges
- **ISO/IEC 18670 SWHIDs:** intrinsic, decentralised, citation-ready
- **Code Commons:** qualification + provenance layer for AI training
- Born in Europe, worldwide scope, supported by UNESCO, **neutral and non-profit**

Everyone in this room is already using it - transitively.

- **Apertus (EPFL · ETH · CSCS)** — Stack v1.2 + CommonPile Stack v2 Edu
- **IFM / K2 / K2-V2 (MBZUAI · Abu Dhabi · Paris)**
- **BharatGen / Param2 (India · IndiaAI Mission)**
- **EleutherAI / The Stack v2 subset**

Stop wasting resources use SWHID!

Join the SCAI Software Heritage

TODAY!

Contribute

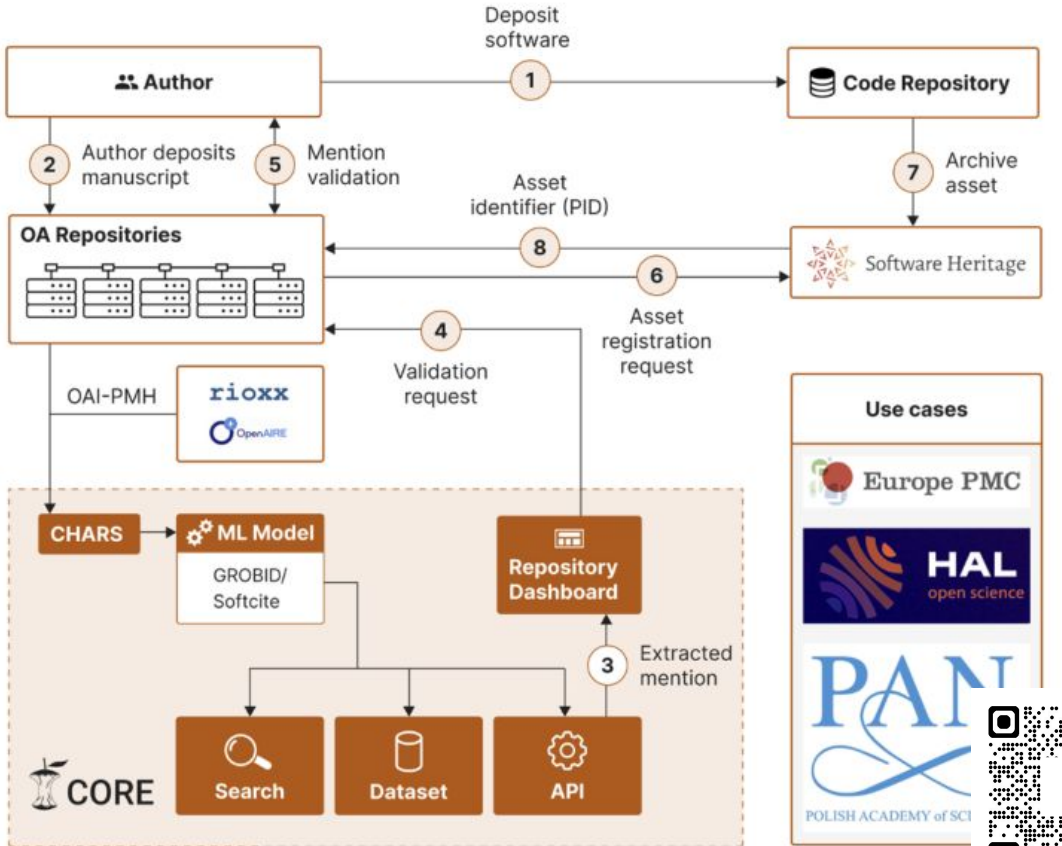
to

CodeCommons

Appendix

Related projects feeding the archive expansion

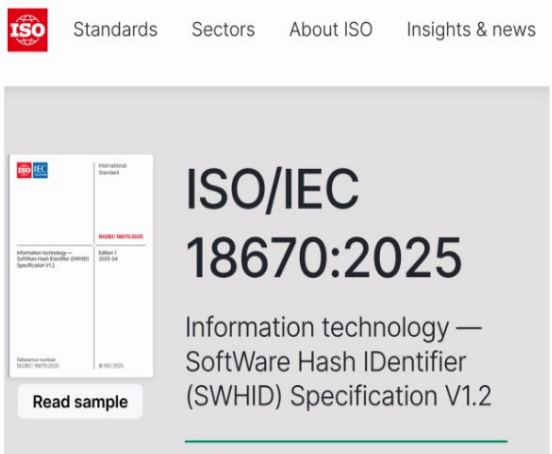
SoFAIR



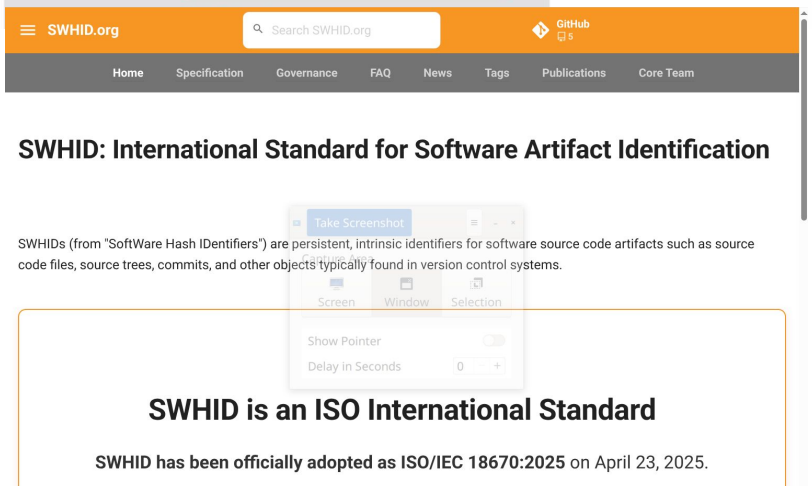
SWH-Security

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