CODE COMMONS

The next generation infrastructure for massive analysis of software source code

bpifrance | SERVIR L'AVENIR

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

GENERATIVE AI FOR CODE : THE OPEN ISSUES

Gousios et al. GHTorrent : GitHub's data from a firehose, MSR 2012

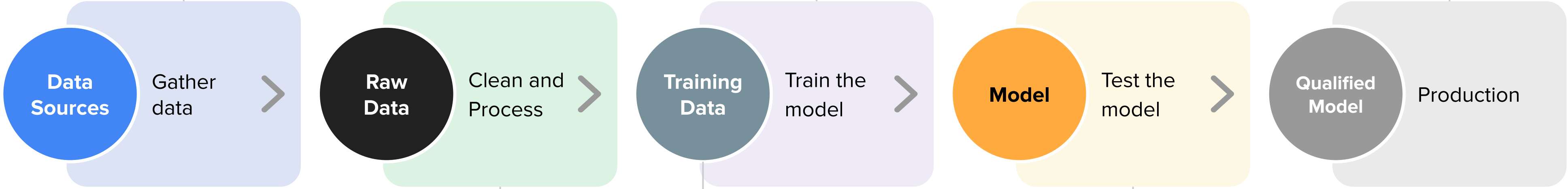
Collect source code, issues, PR, discussions, etc. **is very expensive**. **Redoing** it over and over again **is an anti-ecological waste**.

Lefeuvre et al. Fingerprinting and Building Large Reproducible Datasets REP'23

No precise identification and **lack of availability** of training data are huge obstacles to **transparency** and **reproducibility**.

Sallam et al.
ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns 2023

Lack of **traceability** of generative AI outputs make it **irrespective of authors**



Data Sources — Gather data

Raw Data — Clean and Process

Training Data — Train the model

Model — Test the model

Qualified Model — Production

Building a quality training set is a **very complex task,** redoing it over and over again behind closed doors is a waste of energy and human resources

Gunasekar et al. « Textbooks Are All You Need » 2023
https://arxiv.org/abs/2306.11644

Extracting qualified subsets for training is **difficult** and time consuming.

Ledivarec et al.
HyperDiff: Computing Source Code Diffs at Scale
ASE 2023

Extracting **quality subsets** should allow to **specialize** LLMs to perform **quality programming and software engineering tasks**.

Fan et al. Large language models for software engineering: Survey and open problems
FoSE 2023

□ Issues

Roberto Di Cosmo, 2024

The Opportunity

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Inria avec unesco

Largest archive of open source code

A unique digital commons built since 2015

Cultural Heritage

Industry

Research

Public Administration

Source files

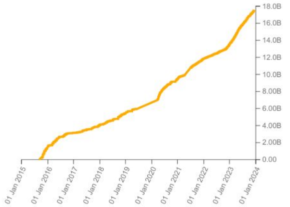
17,567,724,625

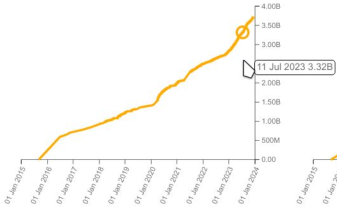
Commits

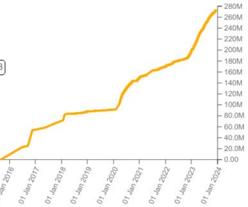
3,730,352,827

Projects

274,163,348







Directories

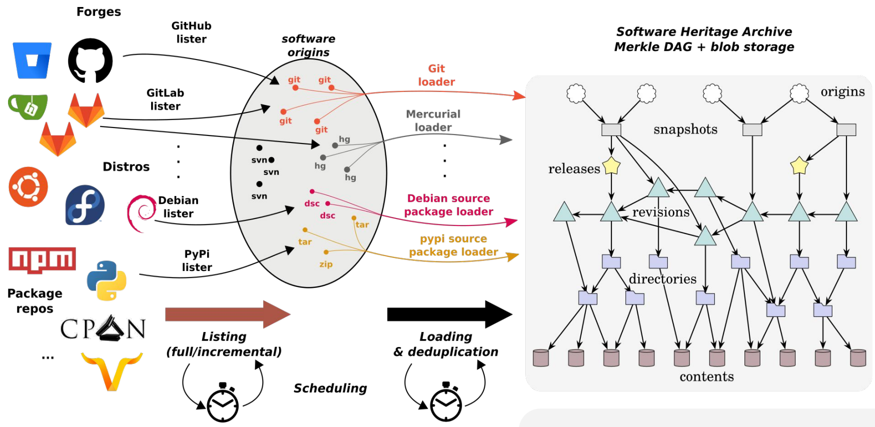
14,101,884,066

Authors

68,895,808

Releases

81,171,066



500+ platforms

All versions, all development history in a single graph

- **50** × **10^9** nodes
- **700** × **10^9** edges

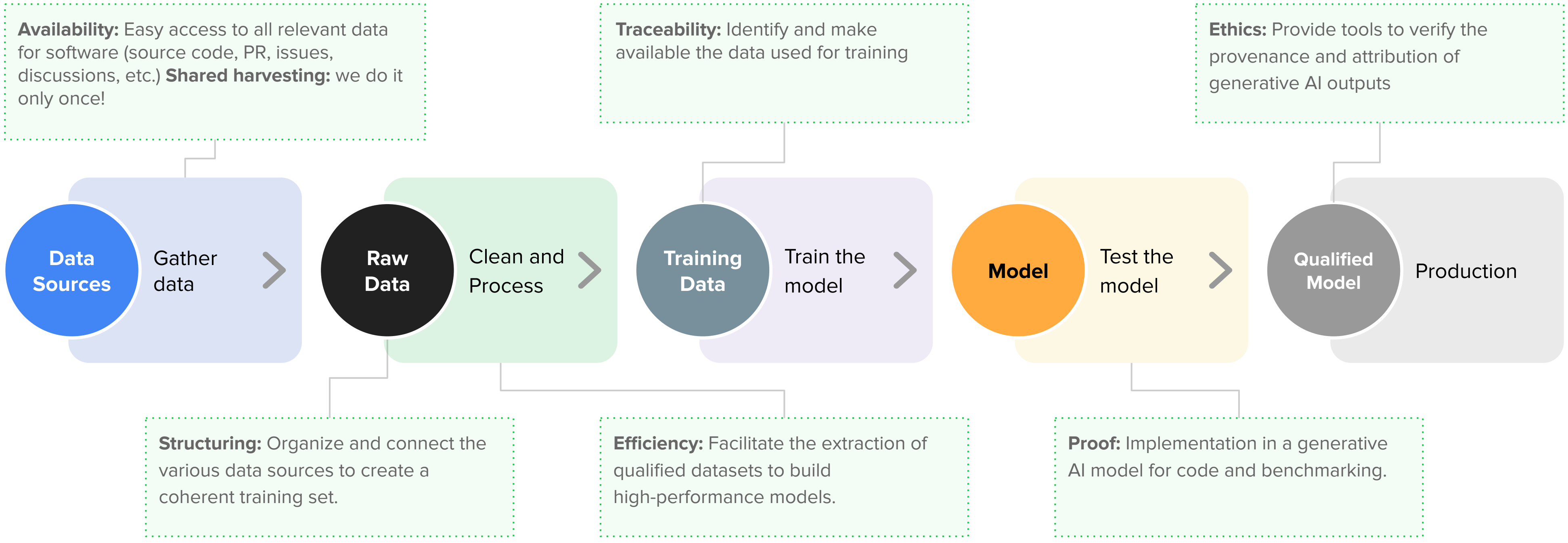
~ **2 Po** of data

ensures **availability**
guarantees **integrity**
allows **traçability**

} Of source codes

A **unique infrastructure**

A STEP FORWARD: 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

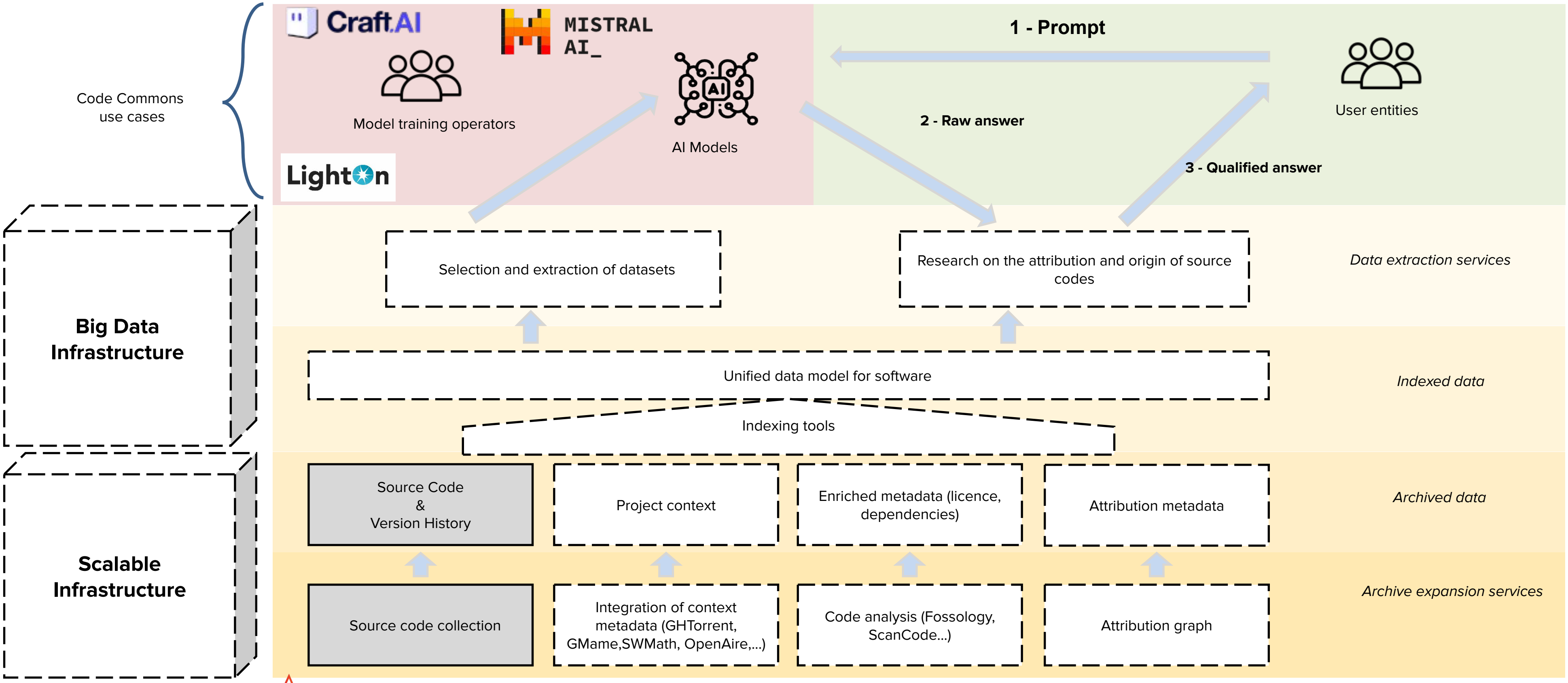
Structuring: Organize and connect the various data sources to create a coherent training set.

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

Proof: Implementation in a generative AI model for code and benchmarking.

🟩 Solutions

CODE COMMONS: bird's eye view



Code Commons use cases

Craft.AI

MISTRAL AI_

Model training operators

AI Models

LightOn

1 - Prompt

User entities

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

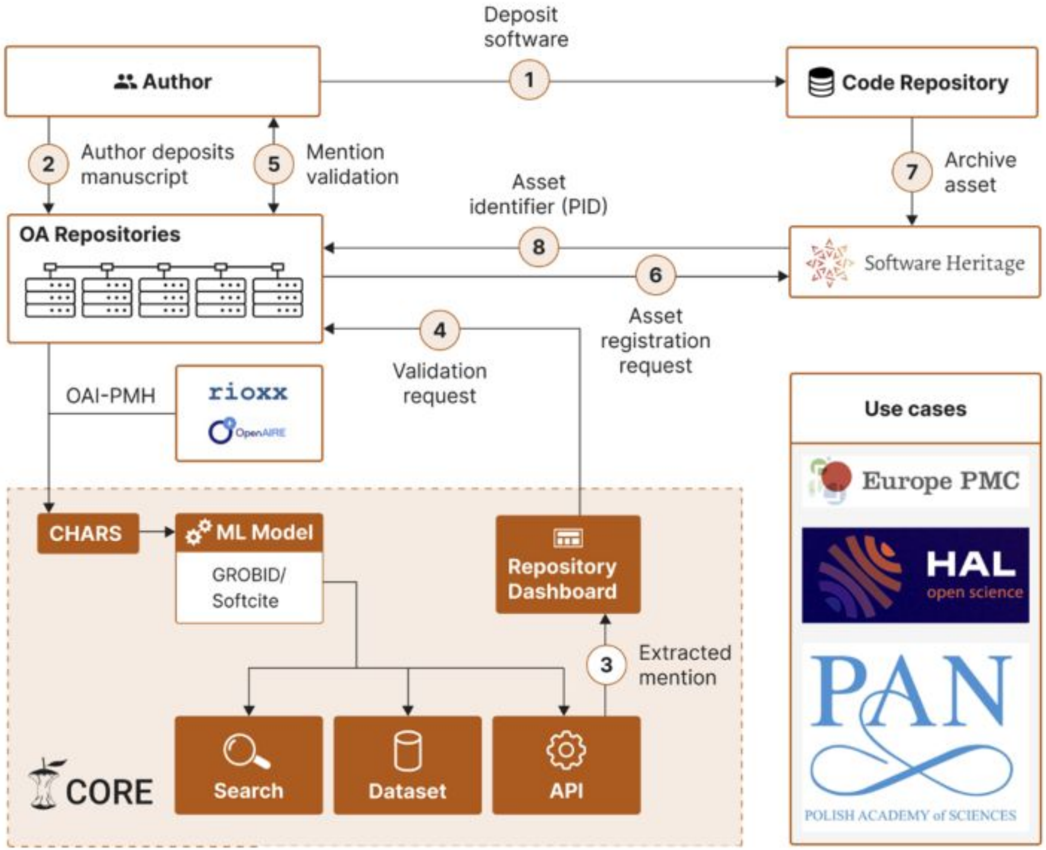
Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

CODE COMMONS : MEET THE TEAMS

Team	Entity / Person	Expertise	
Funded Partners			
Software Heritage	Inria	Universal Software Source Code Archive	
DiverSE Diversity-Centric Software Engineering		Software engineering, code, programming, languages, software variability management Large-scale software evolution, generative AI for software development	
ALMAnaCH		Automatic linguistic modeling and analysis, and computational humanities	
CEDAR		Analysis and processing of large-scale complex data	
DIASI	cea	Natural Language Processing (NLP) Generative AI	
DILS		Engineering, Software, and Systems	
Software Innovation Lab	TWEAG by Modus Create	Machine Learning, Modeling, Natural Language Processing (NLP) Distributed Computing	
Subcontracting (budget < 200k€)			
AboutCode	Philippe Ombredanne	La référence mondiale pour la détection des licences	
External contributors			
Emérite Inria	Patrick Valduriez	Cutting-edge expertise in big data management	
Università di Pisa	Paolo Ferragina Marco Danelutto	Data compression and text algorithms (ACM Paris Kanellakis Award 2022) Expertise in massively parallel programming HPC	
Alma Mater Studiorum Università di Bologna	Maurizio Gabbrielli	Expertise in machine learning and text similarity	
Universita degli Studi di Torino	Marco Aldinucci	EuroHPC and expertise in efficient low-level distributed structures	

Related projects

SoFAIR



SWH-Sec

Clear synergies

- HPC Infrastructure
- Project/code metadata

LLM4Code

“Défi Inria”

- Reliable and productive code assistants based on LLMs
- 10 Inria teams
- Research project