

From Software Heritage to Code Commons

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

Roberto Di Cosmo
Director, Software Heritage
Inria and Université Paris Cité

December 2024



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

- 1 Introduction
- 2 Intermezzo
- 3 Selected highlight: Improving Open Source Security with SWH
- 4 From Software Heritage to CodeCommons
- 5 Conclusion

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- 35+ years of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- 25+ years of Free and Open Source Software
- 15+ years building and directing structures for the common good



1999 *DemoLinux* – first live GNU/Linux distro

2007 *Free Software Thematic Group*

150 members 40 projects 200Me

2008 *Mancoosi project* www.mancoosi.org

2010 *IRILL* www.irill.org

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

2021 *EOSC Task Force on Infrastructures for Software*,
European Union

Software Source Code is Precious Knowledge

Harold Abelson, *Structure and Interpretation of Computer Programs* (1st ed.) 1985
“Programs must be written for people to read, and only incidentally for machines to execute.”

Apollo 11 source code (excerpt)

```
P63SP0T3      CA      BIT6      # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND     CHAN33
              EXTEND
              BZF      P63SP0T4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF      CODE500      # ASTRONAUT:  PLEASE CRANK THE
              TC       BANKCALL     #              SILLY THING AROUND
              CADR     GOPERF1
              TCF      GOTOP00H     # TERMINATE
              TCF      P63SP0T3     # PROCEED   SEE IF HE'S LYING

P63SP0T4      TC       BANKCALL     # ENTER    INITIALIZE LANDING RADAR
              CADR     SETPOS1

              TC       POSTJUMP     # OFF TO   SEE THE WIZARD ...
              CADR     BURNBABY
```

Quake III source code (excerpt)

```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    x2 = number * 0.5F;
    y = number;
    i = * ( long * ) &y; // evil floating point bit level hacking
    i = 0x5f3759df - ( i >> 1 ); // what the fuck?
    y = * ( float * ) &i;
    y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
    // y = y * ( threehalfs - ( x2 * y * y ) ); // 2nd iteration, this
    // can be removed

    return y;
}
```

Len Shustek, Computer History Museum

2006

“Source code provides a view into the mind of the designer.”

Unveiled in 2016



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog



find and reference all software source code

Universal archive



preserve and share all software source code

Research infrastructure



enable analysis of all software source code

A universal software archive, as a shared infrastructure

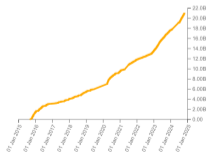
One infrastructure
open and shared



The largest archive ever built

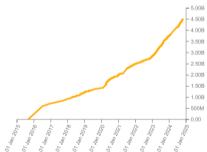
Source files

21,098,377,683



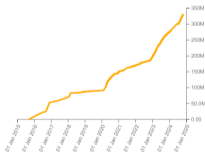
Commits

4,535,490,648



Projects

331,590,016



Directories

16,887,472,535

Authors

83,425,808
Software Heritage

Releases

99,249,418
www.softwareheritage.org

Diamond sponsor



Platinum sponsors



Gold sponsors



Hugging Face

openintervention.com

servicenow



Silver sponsors



Bronze sponsors



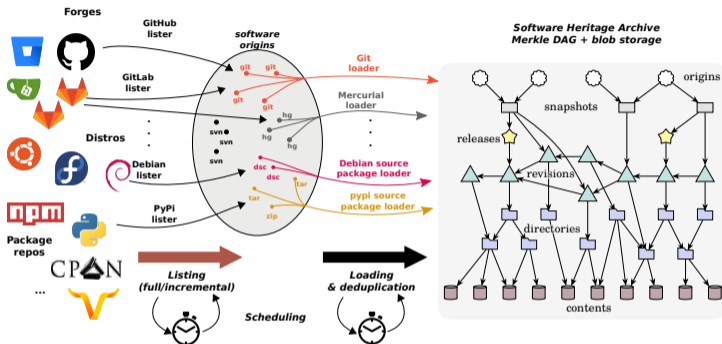
@swheritage

Contact: roberto@dicosmo.org

December 2024

5 / 31

The archive under the hood



Global development history permanently archived in a uniform data model

- over 20 billion unique source files from over 300 million software projects
- ~2PB (compressed) blobs, ~50 B nodes, ~700 B edges

The Software Hash persistent identifier (SWHID)

Software Hash Identifiers (SWHID)

see swhid.org

50+B intrinsic, decentralised, cryptographically strong identifiers, SWHIDs



In [SPDX 2.2](https://spdx.org/specifications); IANA "swh:"; WikiData [P6138](https://www.wikidata.org/wiki/P6138); standardisation ongoing [DIS 18670](https://www.iso.org/standard/718670.html)

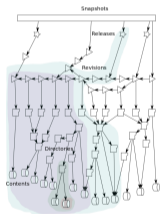
Full fledged *source code references* for traceability, integrity and reproducibility

Examples: [Apollo 11 AGC](https://swhid.org/apollo11), [Quake III rsqrt](https://swhid.org/quake3); Guidelines available: [HOWTO](https://swhid.org/howto) and [ICMS 2020](https://swhid.org/icms2020)

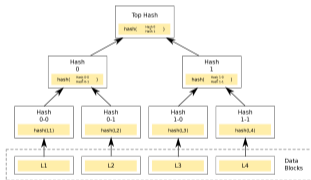
A revolutionary infrastructure

Modern "Library of Alexandria", *international, non profit, long term initiative*
addressing the needs of *industry, research, culture and society as a whole*

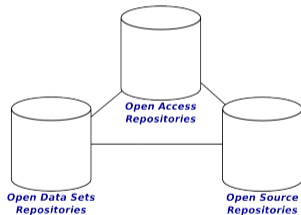
Software Graph



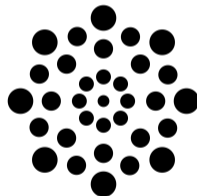
Software Blockchain



Open Science pillar



Big Code



One infrastructure, shared: more efficient, less waste ...
... addressing a broad spectrum of needs!

<https://registry.opendata.aws/software-heritage/>

Registry of Open Data on AWS



Software Heritage Graph Dataset

[digital preservation](#) [free software](#) [open source software](#) [source code](#)

Description

[Software Heritage](#) is the largest existing public archive of software source code and accompanying development history. The Software Heritage Graph Dataset is a fully deduplicated Merkle DAG representation of the Software Heritage archive. The dataset links together file content identifiers, source code directories, Version Control System (VCS) commits tracking evolution over time, up to the full states of VCS repositories as observed by Software Heritage during periodic crawls. The dataset's contents come from major development forges (including GitHub and GitLab), FOSS distributions (e.g., Debian), and language-specific package managers (e.g., PyPI). Crawling information is also included, providing timestamps about when and where all archived source code artifacts have been observed in the wild.

Update Frequency

Data is updated yearly

License

Creative Commons Attribution 4.0 International. By accessing the dataset, you agree with the Software Heritage [Ethical Charter for using the archive data](#) and the [terms of use for bulk access](#).

Documentation

<https://docs.softwareheritage.org/devel/swh-dataset/graph/athena.html>

Managed By

Software Heritage

See all datasets managed by [Software Heritage](#).

Contact

Software Heritage

www.softwareheritage.org

Resources on AWS

Description

Software Heritage Graph Dataset

Resource type

S3 Bucket

Amazon Resource Name (ARN)

```
arn:aws:s3::softwareheritage
```

AWS Region

```
us-east-1
```

AWS CLI Access (No AWS account required)

```
aws s3 ls --no-sign-request s3://softwareheritage/
```

Description

S3 Inventory files

Resource type

S3 Bucket

Amazon Resource Name (ARN)

```
arn:aws:s3::softwareheritage-inventory
```

AWS Region

```
us-east-1
```

AWS CLI Access (No AWS account required)

```
aws s3 ls --no-sign-request s3://softwareheritage-inventory
```

[@swheritage](https://twitter.com/swheritage)

Contact: roberto@dicosmo.org

December 2024

9 / 31

Example: most popular commit verbs (stemmed)

Query using Amazon Athena

```
SELECT COUNT(*) AS C, word FROM (  
  SELECT word_stem(lower(split_part(  
    trim(from_utf8(message)), ' ', 1)))  
  AS word FROM revision  
  WHERE length(message) < 1000000)  
WHERE word != ''  
GROUP BY word  
ORDER BY C  
DESC LIMIT 20;
```

Total cost: approximately .5 euros

Results

Completed Time in queue: 272 ms Run time: 33.545 sec Data scanned: 94.51 GB

Results (20) [Copy](#) [Download results](#)

< 1 > ⚙

#	c	word
1	271573294	updat
2	163328012	merg
3	140044381	add
4	105800317	fix
5	103646653	ad
6	52891401	bump
7	50067041	initi
8	45609622	creat
9	42633225	remov
10	32230842	chang
11	23110410	delet
12	20734745	new
13	16644508	commit
14	15651821	test

State-of-the-art graph compression from social networks



Paolo Boldi, Antoine Pietri, Sebastiano Vigna, Stefano Zacchirolì

Ultra-Large-Scale Repository Analysis via Graph Compression

SANER 2020, 27th Intl. Conf. on Software Analysis, Evolution and Reengineering. IEEE

Results

Full graph structure (50 B nodes, 700 B edges) in 300 GiB RAM

- traversal time is tens of ns per edge
- bidirectional traversals implemented
- **beware:** metadata access is still *off RAM*

Java, gRPC and Rust APIs available

docs.softwareheritage.org/devel/swh-graph/grpc-api.html

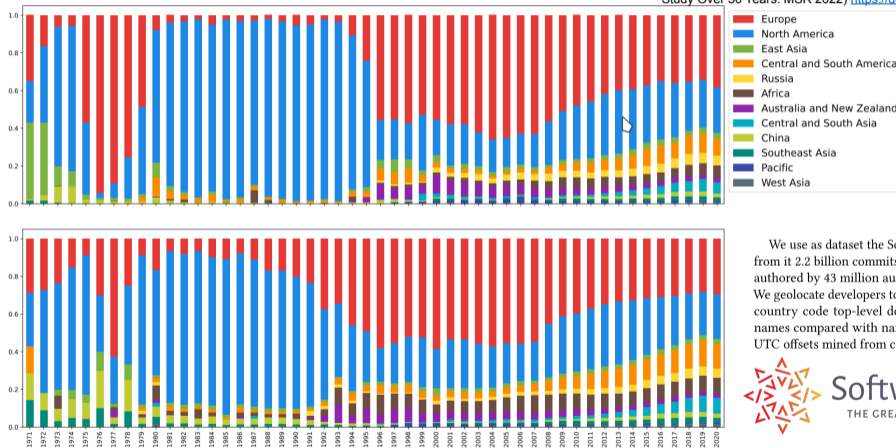
- 1 Introduction
- 2 **Intermezzo**
- 3 Selected highlight: Improving Open Source Security with SWH
- 4 From Software Heritage to CodeCommons
- 5 Conclusion

Because software is naturally international !

MSR '22, May 23–24, 2022, Pittsburgh, PA, USA

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>



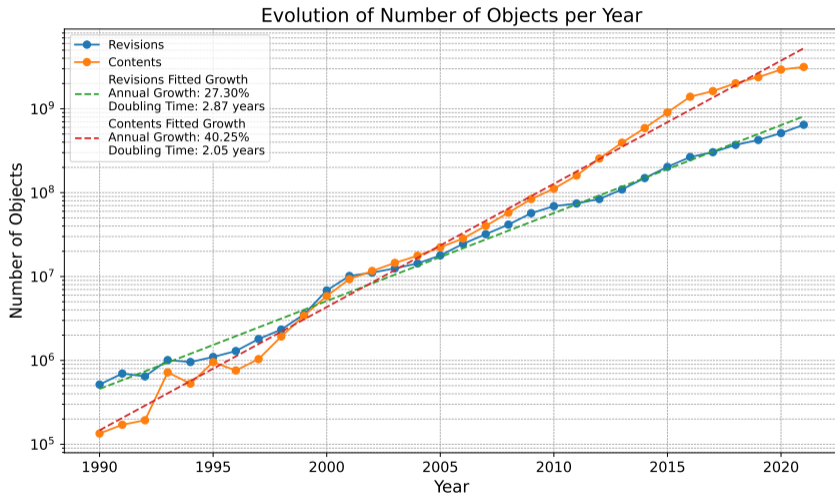
We use as dataset the Software Heritage archive [3] and analyze from it 2.2 billion commits archived from 160 million projects and authored by 43 million authors during the 1971–2021 time period. We geolocate developers to 12 world regions, using as signals email country code top-level domains (ccTLDs) and author (first/last) names compared with name distributions around the world, and UTC offsets mined from commit metadata.



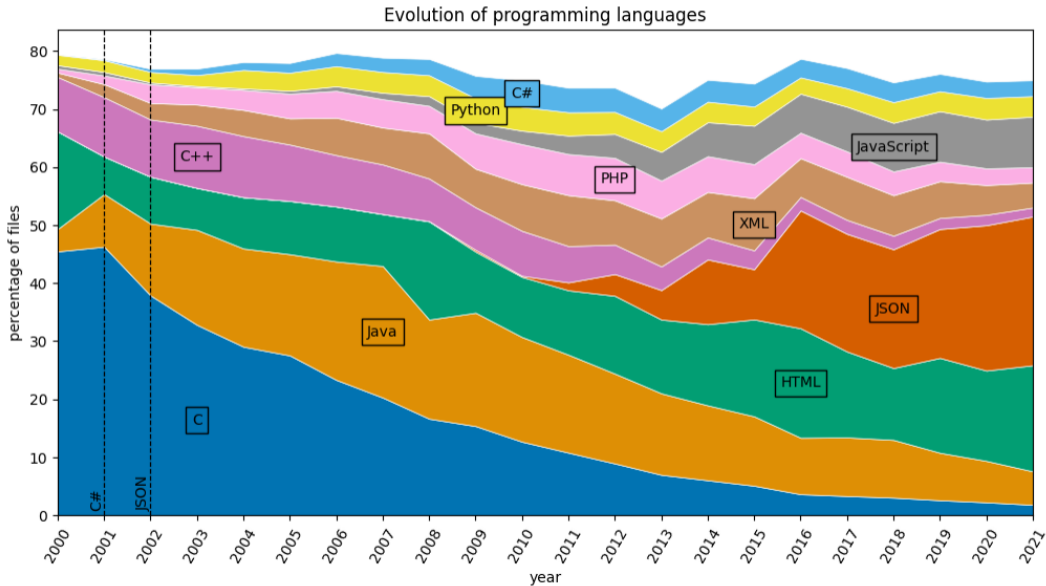
Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Figure 3: Ratio of commits (above) and active authors (below) by world zone over the 1971–2020 period.

30 years of growth of public source code



Programming language evolution over 50 years



- 1 Introduction
- 2 Intermezzo
- 3 Selected highlight: Improving Open Source Security with SWH**
- 4 From Software Heritage to CodeCommons
- 5 Conclusion

CRA

- security obligations for products with digital content put on the market in Europe
- Council vote: 10/10/2024; reporting obligations: ~Q3 2026, full compliance: ~Q4 2027

What Software Heritage brings to the table for Open Source

- long term **availability** (archive)
- **integrity** guarantee (SWHID)
- **traceability** (SWH graph)
- and much more

Breaking news

SWH joins the [Open Regulatory Compliance Working Group](#)



An universal knowledge base about public code vulnerabilities

Vision

- **Software Heritage** is the perfect (and only) place where to build an universal knowledge base that maps known vulnerabilities to public code artifacts.
- We can provide an **open data API mapping SWHIDs to CVEs**, that knows about *all public code commits* and can be leveraged to increase software security for everybody.

Roadmap

- Current status: working prototype that processes OSV.dev data and use it to "color" the entire SWH commit graph (~4 billion commits) with vulnerability information.
- Upcoming feature of the archive. (See: swhsec.github.io)

- 1 Introduction
- 2 Intermezzo
- 3 Selected highlight: Improving Open Source Security with SWH
- 4 From Software Heritage to CodeCommons
- 5 Conclusion

Large Language Models for Code



Image created with DALL-E

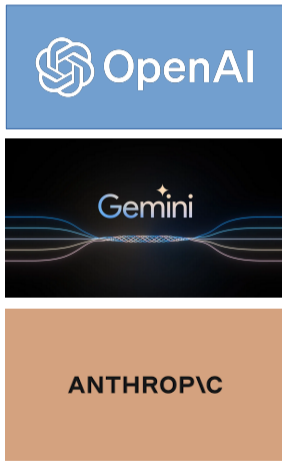
Software source code is massively used for building Large Language Models.

Independently of what we do, there is no turning back.

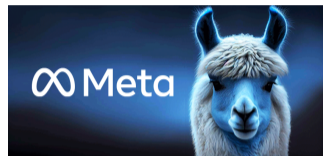
The **real question** is *how* they should be built and *whom* they should benefit.

Let's have a candid look around us.

Closed model APIs



Open model weights



Mistral AI



Closed model APIs

 Model weights not available

- Can't run the model locally
- Can't inspect the model's representations
- Limits fine-tuning abilities

And more:

- limits user freedom
(personal data leakage)

Open model weights

Closed model APIs

Model weights not available

- Can't run the model locally
- Can't inspect the model's representations
- Limits fine-tuning abilities

And more:

- limits user freedom
(personal data leakage)

Open model weights

Training data is not disclosed

- Content 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

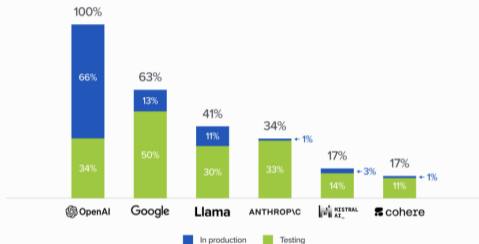
This is not what "open" should mean.

Can we change all this? How?

A window of opportunity: market

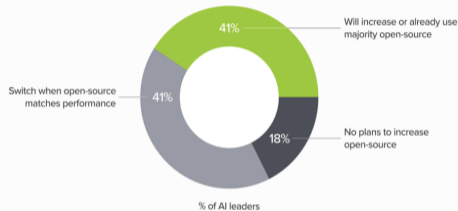
Which model providers are enterprises using?

alo Growth



Enterprise expectations for open source usage in 2024 and onward

alo Growth



LLMs follow a winners take all dynamics...

... but companies want “open source” models

A window of opportunity: regulations

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

Art. 53: 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.



Software Heritage in this picture

Looking for founding principles at Software Heritage

© October 19, 2023

Software Heritage Statement on Large Language Models for Code

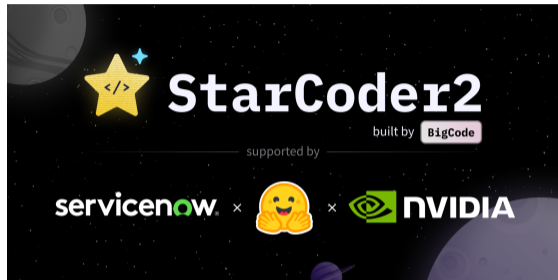
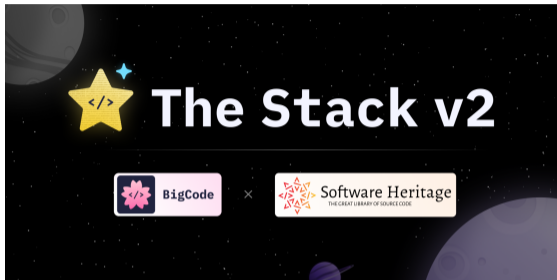


Principles

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.

Question: are we asking too much?

Findings from [BigCode: The Stack v2](#) and [StarCoder2](#)

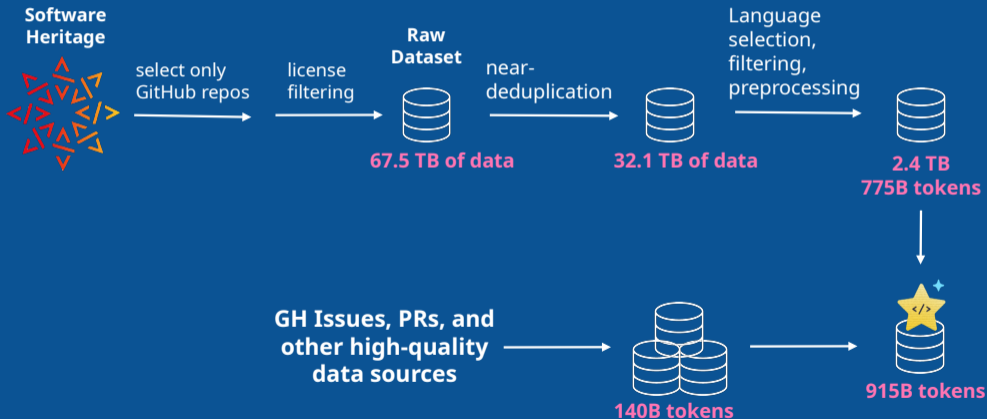


Released February 28th 2024

Yes one can build [the best open LLM for code available](#) while fully adhering to the Software Heritage principles for responsible LLMs, ...
and even more: the full training pipeline is made public too!

The Stack v2

Data collection pipeline fully open and transparent built by BigCode



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 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.

Transparency is easy: [SWHID](#) (undergoing ISO standardisation) and Software Heritage
N.B. : may be mandated by regulations!

Opt out is complex: who is *the real right owner*?
(similar issues to license compliance)



- **Building the training set is complex:** e.g. includes **license compliance** alike work **at massive scale**
- **Generating attribution information on model output is more complex** than license compliance

We need a **coordinated effort** to ensure fully open models will succeed!



CODE COMMONS



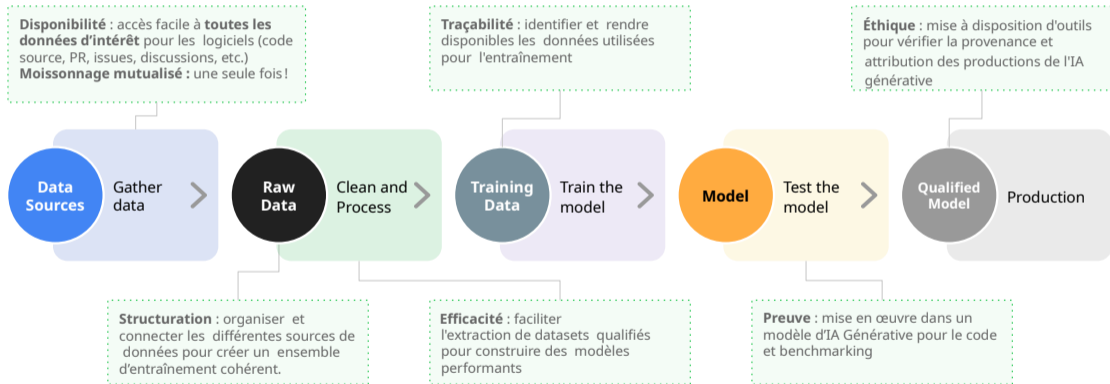
Software Heritage
THE GREAT LIBRARY OF SOURCE CODE



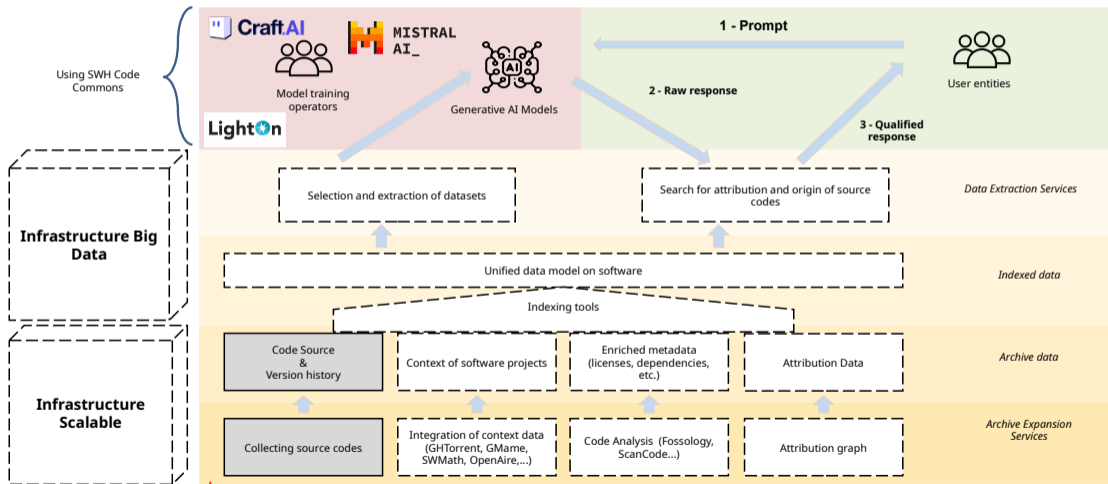
TWEAG
by Mozilla Create

bpifrance | SERVIR CAVENIR

CODE COMMONS : « LA » RÉPONSE



CODE COMMONS : TECHNICAL ARCHITECTURE



CODE COMMONS: LEGAL & ETHICAL ASPECTS

Key questions:

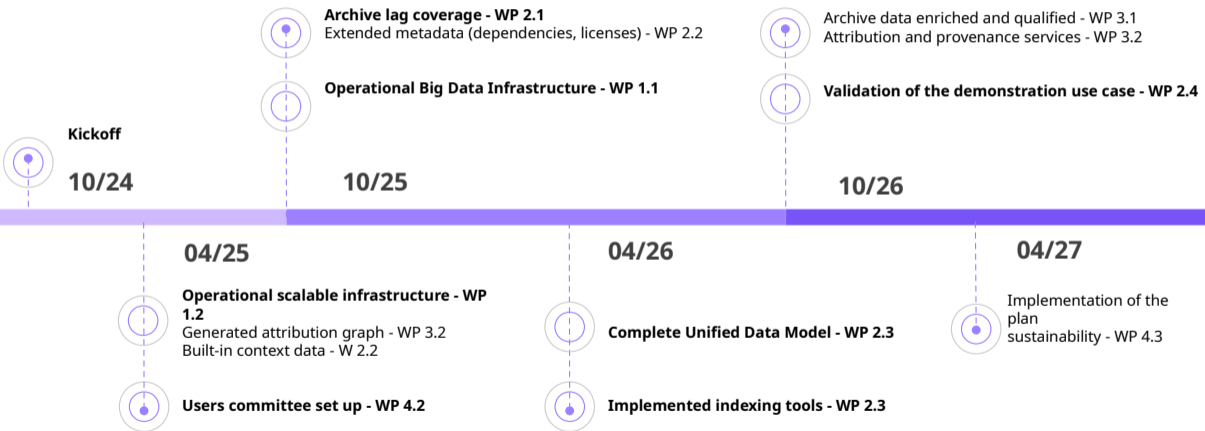
- + Legal framework for *training* models on code
 - + US: Fair Use 4 step test
 - + EU: TDM exception in EU Copyright Directive, Art. 3 (research) and 4 (general)
- + Legal framework for *using* models

Challenges:













- + Provide useful datasets for training, as openly as possible...
- + ... while ensuring respect of key principles...
- + ... and establishing the correct contractual framework

We need a working group focused on these issues

UPDATED EXECUTION SCHEDULE



COMMONS CODE: THE ACTORS

Team	Entity / Referent	Expertise
Funded partners		
 Software Heritage		Universal Archive of Software Source Code
 DiverSE <small>Software Variability Expertise</small>		Software engineering, code, programming, languages, Software variability management Large-scale software evolution Generative AI for software development
 Almanac		Automatic linguistic modeling and analysis and computational humanities
 CEDAR		Analysis and processing of complex, large-scale data
DIASI		Automatic language processing Generative AI
DILS		Engineering, Software and Systems
Software Innovation Lab		Machine learning, Modeling, Natural language processing Distributed computing
	 <small>by Modus Create</small>	
Subcontracting (budget < 200k€)		
 AboutCode	Philippe Ombredanne	The global benchmark for license detection
Unfunded partners		
Emeritus Inria	Patrick Valduriez	Cutting-edge expertise in big data management
 Sant'Anna <small>School of Advanced Studies - Pisa</small>	Paolo Ferragina	Data compression and text algorithms (ACM Paris Kanellakis award 2022)
 UNIVERSITÀ DI PISA	Marco Danelutto	Massively parallel HPC programming expertise
 ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA	Maurizio Gabrielli	Expertise in machine learning and text similarity
 UNIVERSITÀ DEGLI STUDI DI TORINO	Marco Aldinucci	EuroHPC and efficient low-level distributed structure expertise

- 1 Introduction
- 2 Intermezzo
- 3 Selected highlight: Improving Open Source Security with SWH
- 4 From Software Heritage to CodeCommons
- 5 Conclusion

Software Heritage is

- vendor neutral, open source
- worldwide, long term

Software Heritage enables

- archival, reference, integrity
- traceability, global knowledge base

Call to action

- support a shared open infrastructure to support your use cases
- develop new applications, tackle new scientific challenges
- positions open for CodeCommons

Join us



Annual report 2023

