Free/Open Source Software in the Era of AI Hype

Stefano Zacchiroli

2025-03-26

About the speaker

- Professor of Computer Science, Télécom Paris
- Free/Open Source Software activist (25+ years)
- Debian Developer & Former 3x Debian Project Leader
- Former Open Source Initiative (OSI) director
- Software Heritage co-founder & Chief Scientific Officer

In this talk

Digital Commons

Software Heritage

“Open” AI systems

Open Source AI Definition

Appendix

Digital Commons

Software Freedom 101

Definition (Free Software)

A program is **free software** if the program's users have the four **essential freedoms**:

- Freedom #0, to *run* the program, for any purpose
- Freedom #1, to *study* how the program works, and change it
- Freedom #2, to *redistribute* copies
- Freedom #3, to *improve* the program, and *release* improvements

— GNU Project (1986)

[1] License list by the Free Software Foundation (FSF): https://www.gnu.org/licenses/license-list.html

[2] License list by the Open Source Initiative (OSI): https://opensource.org/licenses

Software Freedom 101

Definition (Free Software)

A program is **free software** if the program's users have the four **essential freedoms**:

- Freedom #0, to *run* the program, for any purpose
- Freedom #1, to *study* how the program works, and change it
- Freedom #2, to *redistribute* copies
- Freedom #3, to *improve* the program, and *release* improvements

— GNU Project (1986)

- The definition applies to a "program" (it will be important later).
- Practical implementation: release your code under a free software[1]/open source[2] license.
- Very important political notion in the balance of power: who controls the code, controls its users.

[1] License list by the Free Software Foundation (FSF): https://www.gnu.org/licenses/license-list.html

[2] License list by the Open Source Initiative (OSI): https://opensource.org/licenses

The Software Commons

Commons (*“Biens communs”*)

The **commons** is the cultural and natural resources accessible to all members of a society, including natural materials such as air, water, and a habitable earth. These resources are held in common, not owned privately.

The Software Commons

Commons (*"Biens communs"*)

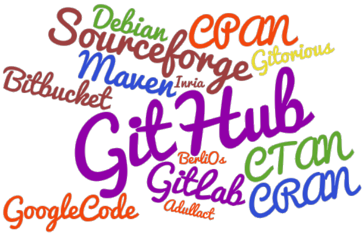
The **commons** is the cultural and natural resources accessible to all members of a society, including natural materials such as air, water, and a habitable earth. These resources are held in common, not owned privately.

Software Commons (*"Bien commun logiciel"*)

The **software commons** consists of all computer software which is available at little or no cost and which can be altered and reused with few restrictions. Thus *all open source software and all free software are part of the [software] commons*.

Every time you publish a new line of code/release a new project under a free/open source software license, you are contributing to **grow the software commons**, available to all.

But *where* is this commons?



Fashion victims

- Many disparate development platforms
- A myriad places where distribution may happen
- Projects tend to migrate from one place to the other over time

But *where* is this commons?

Sourceforge
Debian
CPAN
Gitorious
Maven
Inria
Bitbucket
GitHub
BerliOs
CTAN
Gitlab
CRAN
GoogleCode
Adullact

Fashion victims

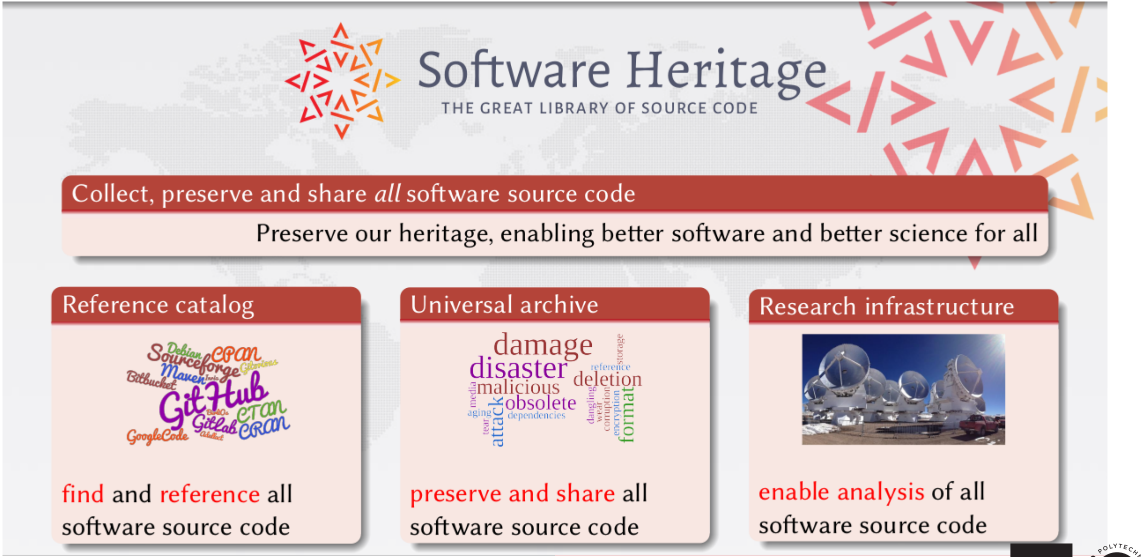
- Many disparate development platforms
- A myriad places where distribution may happen
- Projects tend to migrate from one place to the other over time

One place

… where can we find, track and search *all* source code?

Software Heritage

Software Heritage in a Nutshell



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

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

Reference catalog

find and reference all software source code

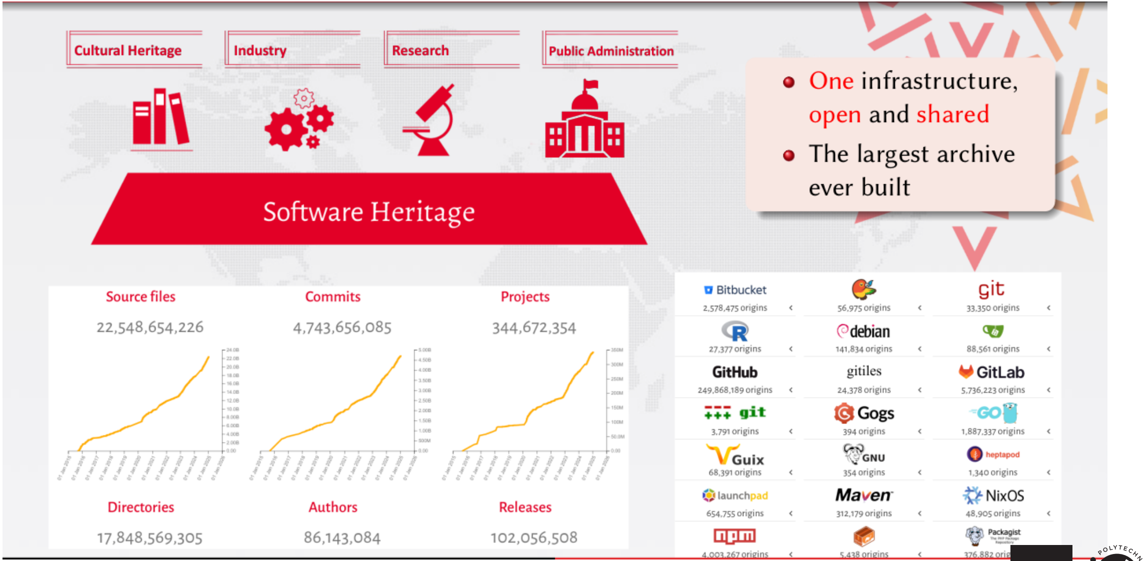
Universal archive

preserve and share all software source code

Research infrastructure

enable analysis of all software source code

A universal software archive, as a shared infrastructure



Cultural Heritage

Industry

Research

Public Administration

Software Heritage

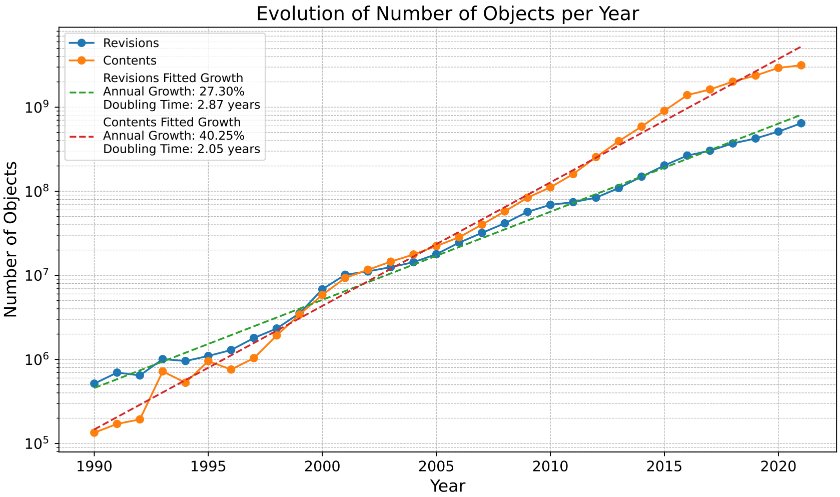
- One infrastructure, open and shared
- The largest archive ever built

Source files	Commits	Projects
22,548,654,226	4,743,656,085	344,672,354
Directories	**Authors**	**Releases**
17,848,569,305	86,143,084	102,056,508

Bitbucket 2,578,475 origins ‹	56,975 origins ‹	git 33,350 origins ‹
R 27,377 origins ‹	debian 141,834 origins ‹	88,561 origins ‹
GitHub 249,868,189 origins ‹	gitiles 24,378 origins ‹	GitLab 5,736,223 origins ‹
git 3,791 origins ‹	Gogs 394 origins ‹	GO 1,887,337 origins ‹
Guix 68,391 origins ‹	GNU 354 origins ‹	heptapod 1,340 origins ‹
launchpad 654,755 origins ‹	Maven 312,179 origins ‹	NixOS 48,905 origins ‹
npm 4,003,267 origins ‹	5,438 origins ‹	Packagist 376,882 orig

Evolution of the Software Commons over Time

… as observable through the lens of Software Heritage:



— Rousseau, Di Cosmo, Zacchiroli. *Software provenance tracking at the scale of public source code*. Empir. Softw. Eng. 25(4): 2930-2959 (2020)

“Open” AI systems

“AI system”

In the AI era, most IT products are no longer just software.

What should we apply the criteria of [software] freedom to?

“AI system”

In the AI era, most IT products are no longer just software.

What should we apply the criteria of [software] freedom to?

Definition (AI system)

An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. […]

— OECD AI Principles (2019)

Components of an “AI system”

But what constitutes an *actual* “AI system”, as used today on the market?[3]

1. Training data ← data of various kinds, through multiple processing stages
2. Training pipeline ← software + data
3. Model weights ← data (float matrices)
4. Inference code ← software + data

[3] This list simplifies things *a lot*, for ease of discussion. A more detailed list can be found in the Model Openness Framework (MOF) https://arxiv.org/abs/2403.13784 (2024).

Components of an “AI system”

But what constitutes an *actual* “AI system”, as used today on the market?[3]

1. Training data ← data of various kinds, through multiple processing stages
2. Training pipeline ← software + data
3. Model weights ← data (float matrices)
4. Inference code ← software + data

Q: What does it mean for an AI system, made of the above components, to be free/open w.r.t. the traditional 4 freedoms?

[3] This list simplifies things *a lot*, for ease of discussion. A more detailed list can be found in the Model Openness Framework (MOF) https://arxiv.org/abs/2403.13784 (2024).

Open Source AI Definition

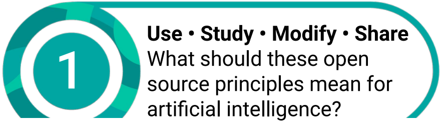
Open Source AI Definition (OSAID) — Timeline

- 2022: The Open Source Initiative (OSI)—the international non-profit organization in charge of maintaining the list of licenses considered to be "open source"—picks up the task of defining **what "open source" means for "AI systems"**.
- ~2 year "co-design process" with multiple stakeholders: activists, academia, civil society, industry.
  - Disclosure: I have participated in the process, as an independent academic.
- October 2024: OSAID (Open Source AI Definition) 1.0 is released[4]

[4] https://opensource.org/ai/open-source-ai-definition

OSAID — Four Freedoms

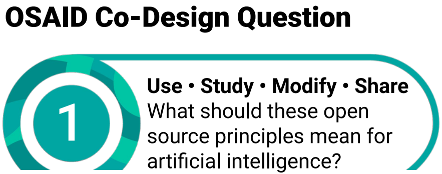
OSAID Co-Design Question



Use • Study • Modify • Share

What should these open source principles mean for artificial intelligence?

OSAID — Four Freedoms

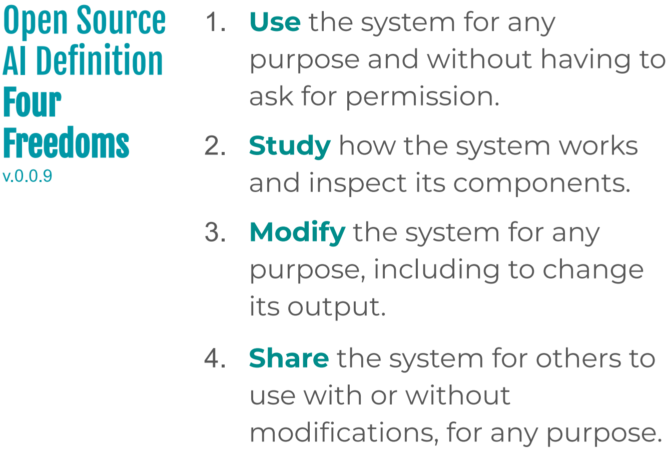


OSAID Co-Design Question

1

Use • Study • Modify • Share
What should these open source principles mean for artificial intelligence?

… pretty straightforward answer, but still unhelpful on practical requirements.



Open Source AI Definition Four Freedoms

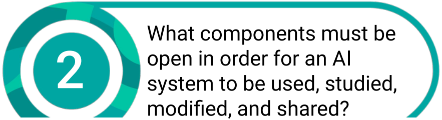
v.0.0.9

1. **Use** the system for any purpose and without having to ask for permission.
2. **Study** how the system works and inspect its components.
3. **Modify** the system for any purpose, including to change its output.
4. **Share** the system for others to use with or without modifications, for any purpose.

— Stefano Maffulli, OSAID townhall 21

OSAID — Required Open Components (1/3)

OSAID Co-Design Question

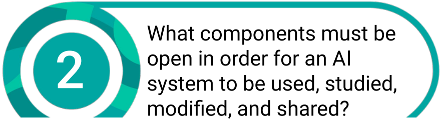


2

What components must be open in order for an AI system to be used, studied, modified, and shared?

OSAID — Required Open Components (1/3)

OSAID Co-Design Question



What components must be open in order for an AI system to be used, studied, modified, and shared?

… here things become more tricky.

What is the **preferred form to make modifications** to machine-learning systems?

The preferred form of making modifications to a machine-learning system must include all the elements below:

Code

- **Code:** The complete source code used to train and run the system. The Code shall represent the full specification of how the data was processed and filtered, and how the training was done. Code shall be made available under OSI-approved licenses.
  - For example […] this must include code used for processing and filtering data, code used for training including arguments and settings used, validation and testing, supporting libraries like tokenizers and hyperparameters search code, inference code, and model architecture.

— OSAID 1.0

OSAID — Required Open Components (2/3)

The preferred form of making modifications to a machine-learning system must include all the elements below:

> **Parameters**
>
> - **Parameters:** The model parameters, such as weights or other configuration settings. Parameters shall be made available under OSI-approved terms.
>   - For example, if used, this might include checkpoints from key intermediate stages of training as well as the final optimizer state.
>
> — OSAID 1.0

OSAID — Required Open Components (3/3)

The preferred form of making modifications to a machine-learning system must include all the elements below:

Data

- **Data Information:** Sufficiently detailed information about the data used to train the system so that a skilled person can build a substantially equivalent system. Data Information shall be made available under OSI-approved terms.

  In particular, this must include:

  1. the complete **description of all data** used for training, including (if used) of **unshareable data**, disclosing the provenance of the data, its scope and characteristics, how the data was obtained and selected, the labeling procedures, and data processing and filtering methodologies;
  2. a listing of all **publicly available training data** and where to obtain it; and
  3. a listing of all training **data obtainable from third parties** and where to obtain it, including for fee.

— OSAID 1.0

OSAID and the AI/Data Conundrum

- The data part of an “AI system” that is OSAID-compliant is **not guaranteed to be open data.**[5]
  - Datasets can be non publicly available and/or licensed under non free/open licenses.
- OSAID has been criticized for this. (Disclosure: including by yours truly.)

[5] In the traditional sense of “data that are openly accessible, exploitable, editable and shareable by anyone for any purpose”, https://okfn.org/en/library/what-is-open/.

OSAID and the AI/Data Conundrum

- The data part of an “AI system” that is OSAID-compliant is **not guaranteed to be open data.**[5]
  - Datasets can be non publicly available and/or licensed under non free/open licenses.
- OSAID has been criticized for this. (Disclosure: including by yours truly.)

Arguments *in favor* of the choice made by OSAID on data include:

- There are important kinds of AI systems on the market, that we want to make “as open as possible” but that will never be able to release their datasets, for legal reasons (e.g., privacy/PII, non-copyrighted data).
  - E.g.,: health-related data, tax information, etc.

[5] In the traditional sense of “data that are openly accessible, exploitable, editable and shareable by anyone for any purpose”, https://okfn.org/en/library/what-is-open/.

Thought Experiment on Free/Open AI Systems

- OSAID is stricter than 1.0 and mandates open data for all datasets.
- There is country where only free/open source software is allowed in the public sector (good!).
- Hence the country only allows to deploy OSAID-compliant AI systems (good!).
- There is an AI system that reliably allows to save lives, but is trained on non-open data.
- You cannot use that system :-(

This is no *proof* that OSAID should *not* be stricter, just evidence that there are difficult ethical choices to be made.

Thought Experiment on Free/Open AI Systems

- OSAID is stricter than 1.0 and mandates open data for all datasets.
- There is country where only free/open source software is allowed in the public sector (good!).
- Hence the country only allows to deploy OSAID-compliant AI systems (good!).
- There is an AI system that reliably allows to save lives, but is trained on non-open data.
- You cannot use that system :-(

This is no *proof* that OSAID should *not* be stricter, just evidence that there are difficult ethical choices to be made.

Noteworthy alternative: use a **“degrees of freedoms”** approach (note: OSI has decided against this). For example, in the Model Openness Framework (MOF):
Class III - *Open Model* < Class II - *Open Tooling Model* < Class I - *Open Science Model*

State of AI Openness on the Market

What’s the current state of freedom/openness of popular AI systems on the market?

State of AI Openness on the Market

What's the current state of freedom/openness of popular AI systems on the market? **It sucks.**

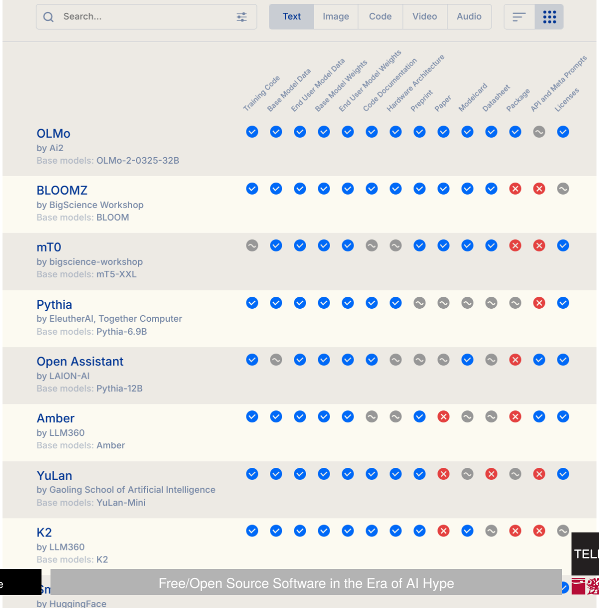
- Many self-declared "open" models only release weights and do so under non free/open license.
  - From Llama license: "If […] the monthly active users of the products […] is greater than 700 million monthly active users in the preceding calendar month, you must request a license from Meta".
  - From RAIL (most popular licenses on Hugging Face): "You agree not to use the Source Code […] 2(b): To generate or disseminate false information with the purpose of Harming others".
- Release of complete training pipelines is rare, but growing.
- Release of complete training datasets is very rare.

European Open Source AI Index

My favorite initiative out there for the **continuing assessment of model openness**: https://osai-index.eu/the-index.

“The European Open-Source AI Index collects information on model openness, licensing, and EU regulation of generative AI systems and providers. The index is a non-profit public resource hosted at the Centre of Language and Speech Technology, Radboud University, The Netherlands, maintained by a small team of academics and community members.”

Liesenfeld, Dingemanse. Rethinking open source generative AI: open-washing and the EU AI Act. ACM FAccT’24. https://doi.org/10.1145/3630106.3659005



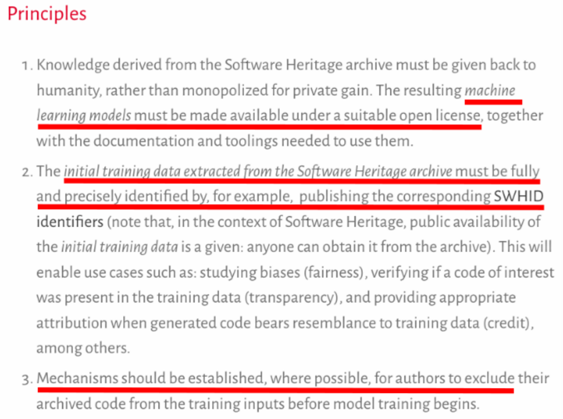
	Training Code	Base Model Data	End User Model Data	Base Model Weights	End User Model Weights	Code Documentation	Hardware Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API and Meta Prompts	Licenses
OLMo by Ai2 Base models: OLMo-2-0325-32B	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	~	✓
BLOOMZ by BigScience Workshop Base models: BLOOM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	~
mT0 by bigscience-workshop Base models: mT5-XXL	~	✓	✓	✓	✓	~	~	✓	✓	✓	✓	✗	✗	✓
Pythia by EleutherAI, Together Computer Base models: Pythia-6.9B	✓	✓	✓	✓	✓	✓	✓	~	~	~	~	~	✗	✓
Open Assistant by LAION-AI Base models: Pythia-12B	✓	~	✓	✓	✓	✓	~	~	~	✓	~	✗	✓	✓
Amber by LLM360 Base models: Amber	✓	✓	✓	✓	✓	~	~	✓	✗	~	~	✗	✓	✓
YuLan by Gaoling School of Artificial Intelligence Base models: YuLan-Mini	✓	✓	✓	✓	✓	✓	✓	✓	✗	~	✗	~	✗	✓
K2 by LLM360 Base models: K2	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	~	✗	✗	~
Sn by HuggingFace														

Software Heritage — AI Principles

Many actors would like to train models on content of the Software Heritage archive. We took a principled approach to decide with whom we can potentially collaborate.

October 19, 2023

Software Heritage Statement on Large Language Models for Code

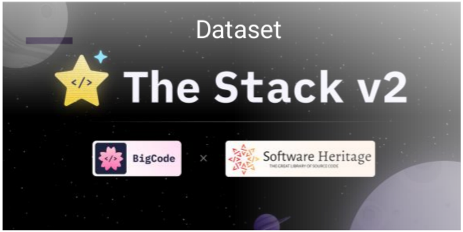


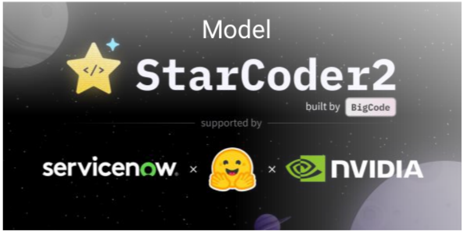
Principles

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.

In the context of a *public* archive like Software Heritage, data is **available by default**, hence it is **identifiability** that becomes the next big requirement to enable reuse and modification.

Findings from BigCode: The Stack V2 and StarCoder2

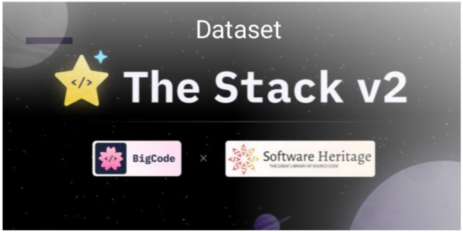


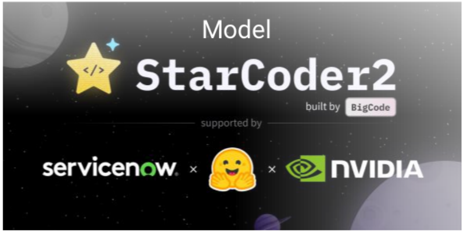


Released February 28th 2024

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!

Findings from BigCode: The Stack V2 and StarCoder2





Released February 28th 2024

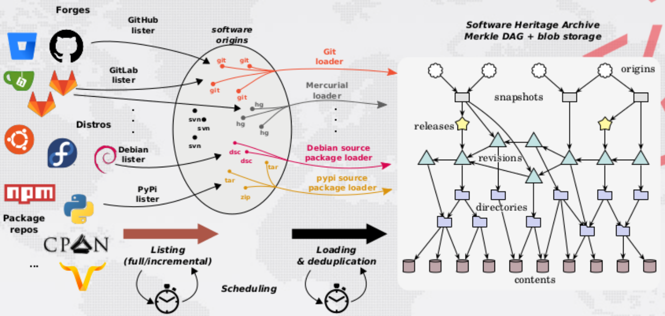
Yes one can build **the best open LLM for code available** while fully adhering to the Software Heritage principles for responsible LLMs, ...
and even more: the full training pipeline is made public too!

Still not completely satisfying:

- Model under RAIL license — It was the go-to-choice at the time for "open" models. Things have improved since, with more and more models under traditional lax/permissive licenses.
- Training dataset includes publicly available but non-explicitly-licensed content, differently from The Stack v1 that only included permissively licensed code.

Appendix

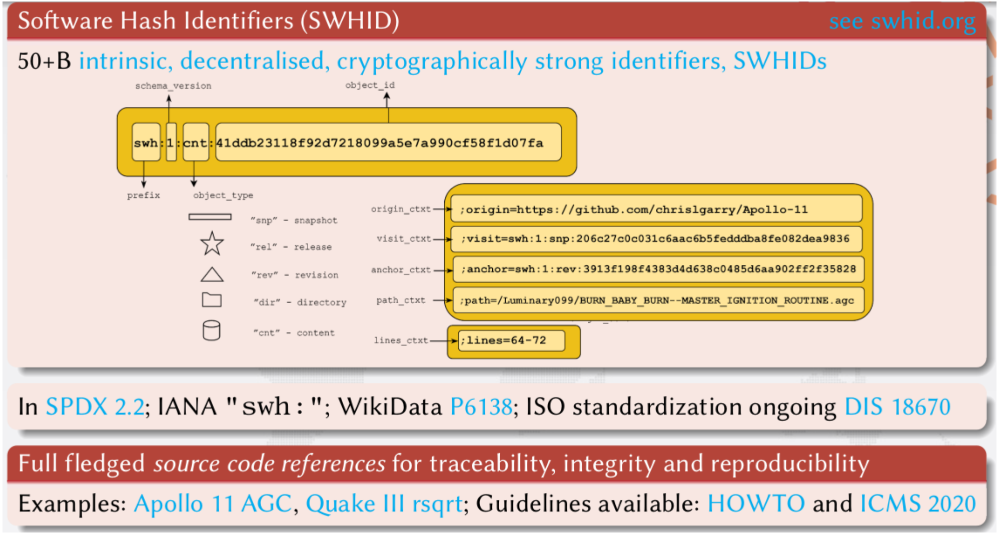
The Software Heritage archive under the hood



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

Software Hash IDentifiers (SWHIDs)



Software Hash Identifiers (SWHID)

see swhid.org

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

```
swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa
;origin=https://github.com/chrislgarry/Apollo-11
;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836
;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828
;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc
;lines=64-72
```

In SPDX 2.2; IANA "swh:"; WikiData P6138; ISO standardization ongoing DIS 18670

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

Examples: Apollo 11 AGC, Quake III rsqrt; Guidelines available: HOWTO and ICMS 2020