Software Heritage

Project presentation and applications to AI

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

We are building the « Library of Alexandria » of software source code

Mission to collect, preserve, and share all software that is publicly available in source code form

A non profit organization

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE
https://www.softwareheritage.org/

Started at Inria in 2016

Inria

Backed by many other organizations

- Public sector
- Academic
- Industry
- ...

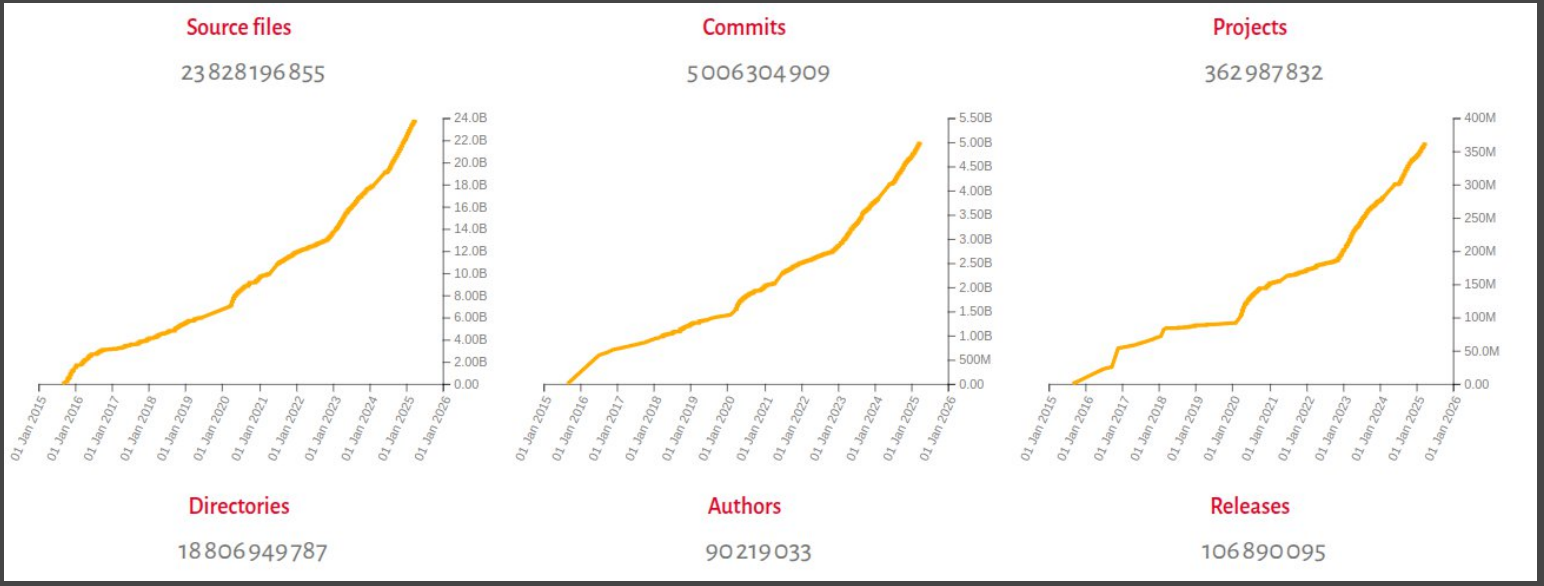
One infrastructure open and shared

The largest archive ever built

Mission: collect, preserve and share all software source code

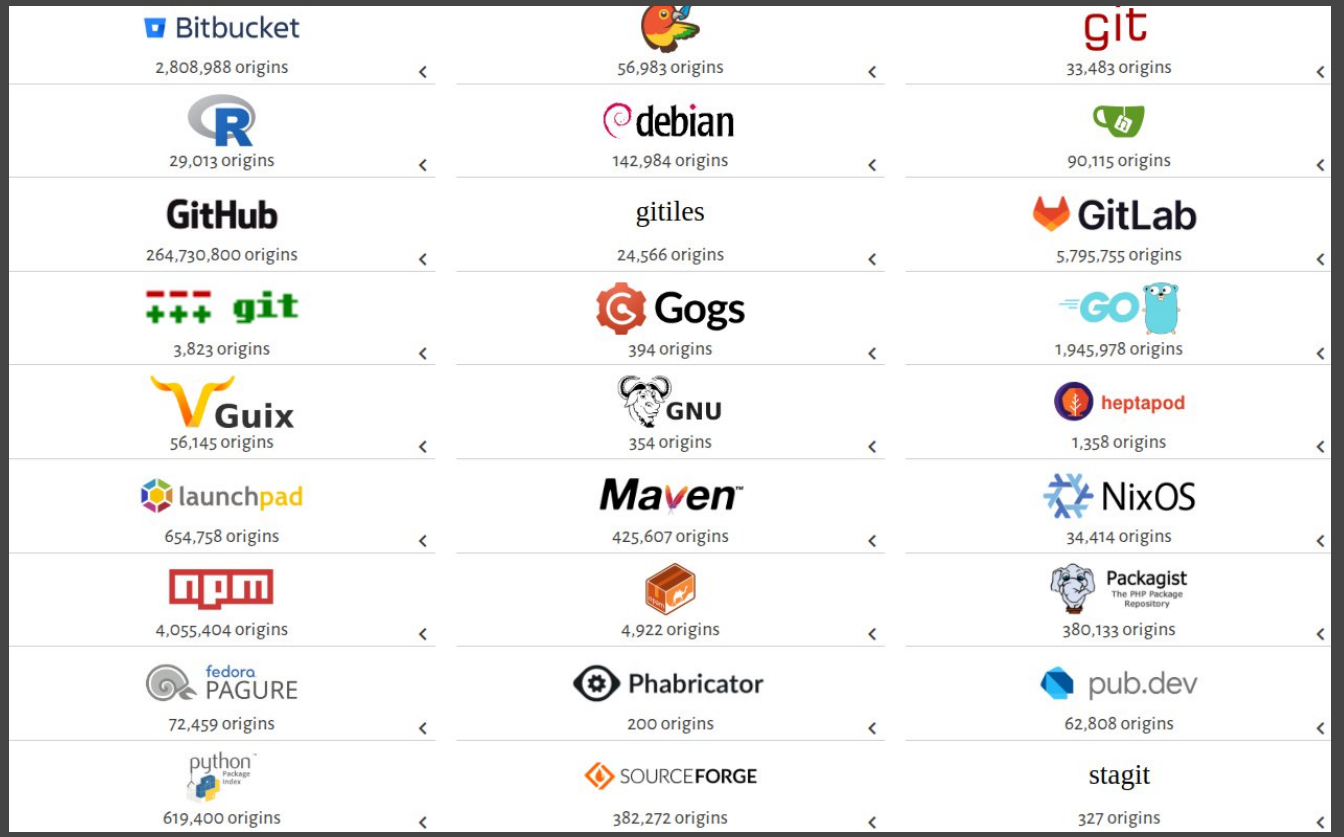
Data we collect

- Source code files: 26B deduplicated source files
- History: 50B of code revisions, with date, author, message, related files...
- Metadata:
  - From the code itself: codemeta.json
  - From the ecosystem around
    - GitHub stars, declared languages
- Releases, specific versions tagged by developers
- Packages
- ...



Data we collect

From a lot of origins:



Some of them do not exist anymore:

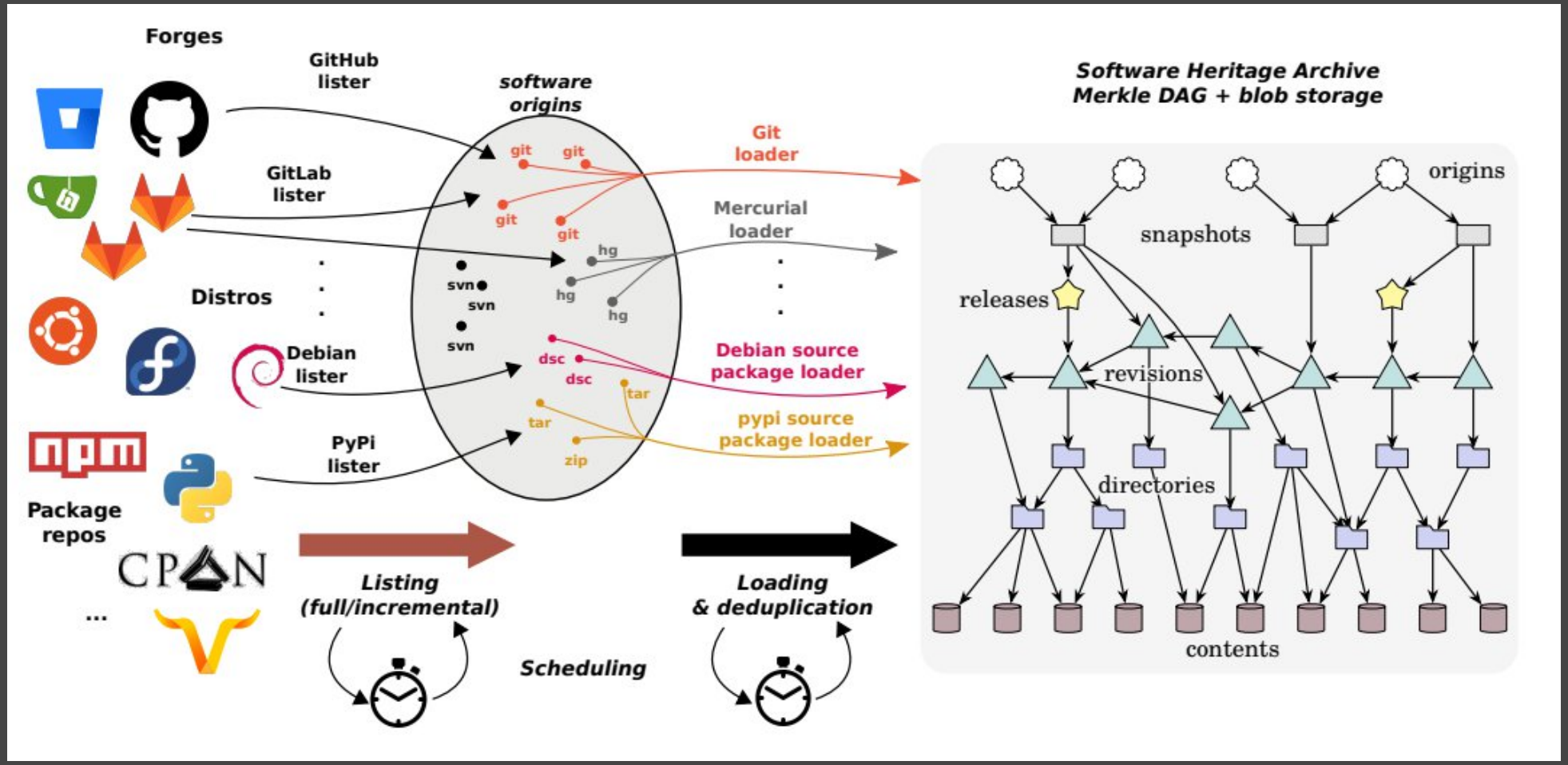
How we archive ?

- Bots “crawling” software forges and translating to our data model
- Deposit, from Libraries advocating for Open Science
- SWHAP: Software Heritage Acquisition Process for older historical source code

- With some technical gems:
  - Object Storage, to store reliably 23B files representing 1.5PB of data
  - Graph compression algorithm, to be able to access to the whole code history from one server
  - SWHID, to identify precisely source code

- https://archive.softwareheritage.org/ if you want to browse the archive content
- https://gitlab.softwareheritage.org/ if you want to see archive code

In a nutshell



- Global development history, permanently archived in a uniform data model
- One infrastructure, shared: more efficient, less waste
- Universal knowledge base for software compliance

Work in progress related to AI

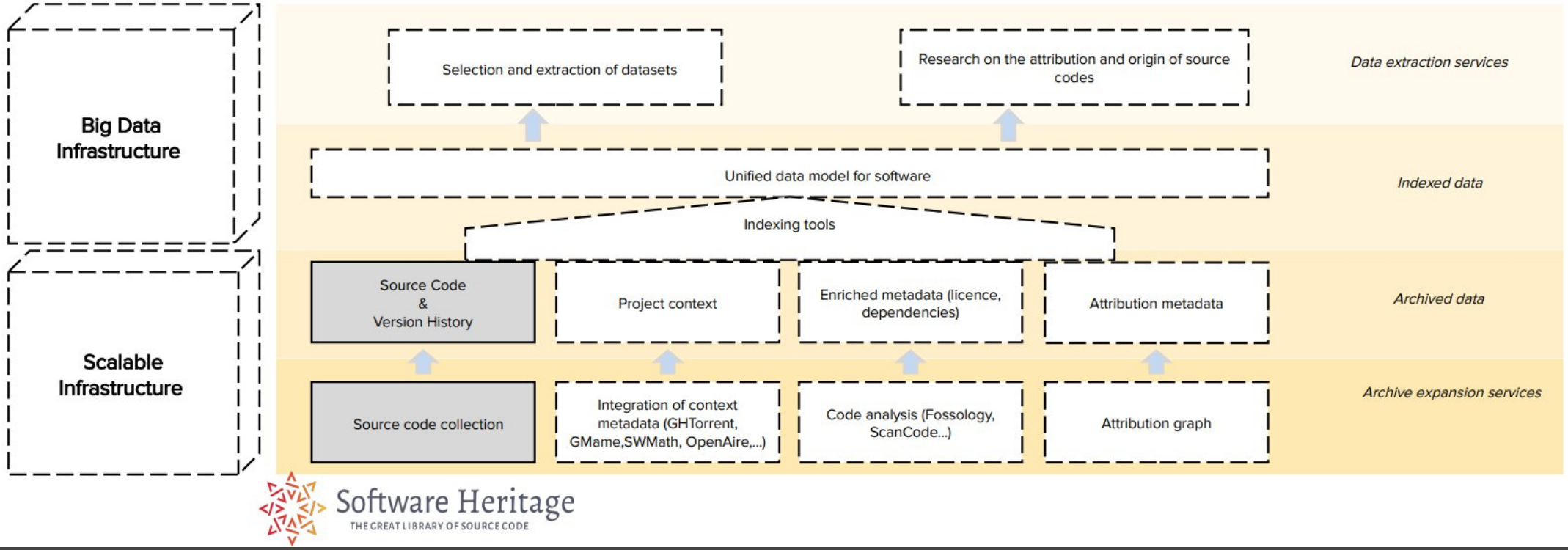
CodeCommons

- Project with many collaborations to:
  - Archive more data, faster
  - Collect more metadata
    - License
    - Language Detection
    - Version and Dependencies Detection
    - Interaction with forges (issues, merge requests, comments...)
  - Detect similar code

- Build LLMs training datasets
  - A self service shop, where you can ask “most important codes in python or golang, updated in the last two years, with this pool of licenses, no known vulnerabilities and used by academics in the field of astronomy”
- Massive, transparent and and shared infrastructure with traceability of training data and code attribution
- Can become some “mini software heritage” to make more science about code

- https://codecommons.org/

CodeCommons

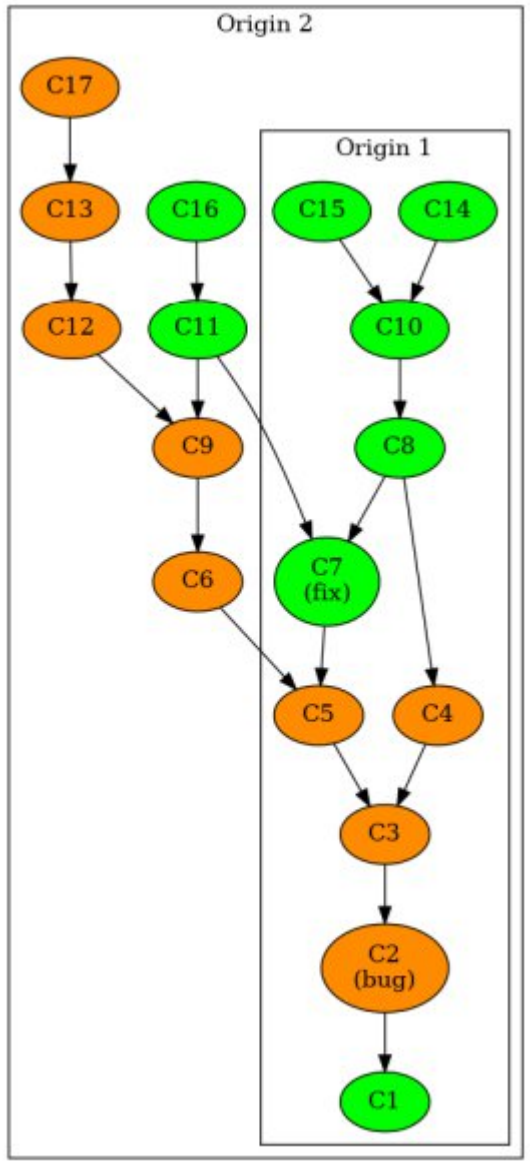


- Project with many collaborations to:
  - Archive more data
  - Collect more metadata
    - License, Language, Versions, dependencies Detection
    - Interaction with forges (issues, merge requests, comments...)
  - Detect similar code

CodeCommons

- Software Heritage statement on LLM for code
  - Open, responsible, and transparent AI: Our shared goal
  - The resulting machine learning models must be made available under a suitable open license
  - The initial training data extracted from the Software Heritage archive must be fully and precisely identified by, for example, publishing the corresponding SWHID identifiers
  - Mechanisms should be established, where possible, for authors to exclude their archived code from the training inputs before model training begins

- Working with AI preferences group
  - Creative Commons / IETF / rslstandard



SWH-SEC / Sec4AI4Sec

- Research work in progress to :
  - Collect CVE
  - Link CVE to files in the history graph
  - Extend the graph with vulnerabilities
- Build datasets of introducing and fixing commits to enable machine learning analysis
  - For instance, diff analysis

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

You can help

- Publish your code in it through HAL/Zenodo/..
- Advocating for Software Heritage
- Joining the communities around Software Heritage
- Becoming an early user, helping us build tool tailored for your need
- Becoming a sponsor, a mirror...

Thank you !

https://www.softwareheritage.org/
https://archive.softwareheritage.org/
https://gitlab.softwareheritage.org/
https://codecommons.org/
https://swhid.org/
https://swhsec.github.io/

thomas.aynaud@inria.fr

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

United Nations
Educational, Scientific and
Cultural Organization

Paris Call

«Software source code represents unique knowledge of humanity's recent history.

It is therefore crucial to work together collectively so that the knowledge embedded in software source code is properly preserved, valued and sharedwith all.

This lies at the core of UNESCO's cooperation with Inria to support the creation of Software Heritage, the global archive of software source code»