SHARING & REUSE

CONFERENCE

OPEN.SHARE.LINK.

Software Heritage: the largest archive of reusable Free and Open Source Software

STEFANO ZACCHIROLI

Co-founder and CTO of Software Heritage project

Follow discussions online @EU_isa2 ISA2 Programme

#SRCONF19

Software Heritage, in a nutshell

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share the *source code* of *all the software*

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

Software Heritage, in a nutshell

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share the *source code* of *all the software*

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

Reference catalog

find and reference all the source code

Software Heritage, in a nutshell

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share the *source code* of *all the software*

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

Reference catalog

find and reference all the source code

Universal archive

preserve all the source code

Software Heritage, in a nutshell

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share the *source code* of *all the software*

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

Reference catalog

find and reference all the source code

Universal archive

preserve all the source code

Research infrastructure

enable analysis of all the source code

Cultural Heritage
Industry
Research
Education
Software Heritage

A principled infrastructure

Technology

- transparency and FOSS
- replicas all the way down

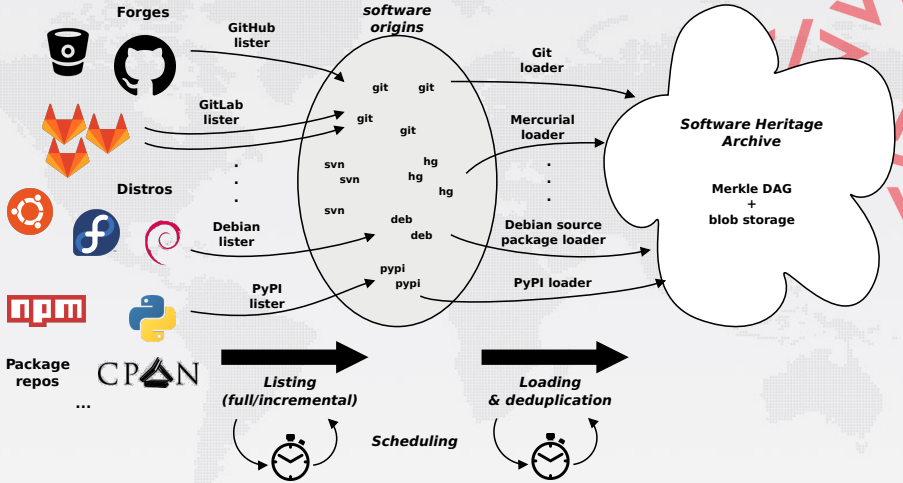
Content

- intrinsic identifiers
- facts and provenance

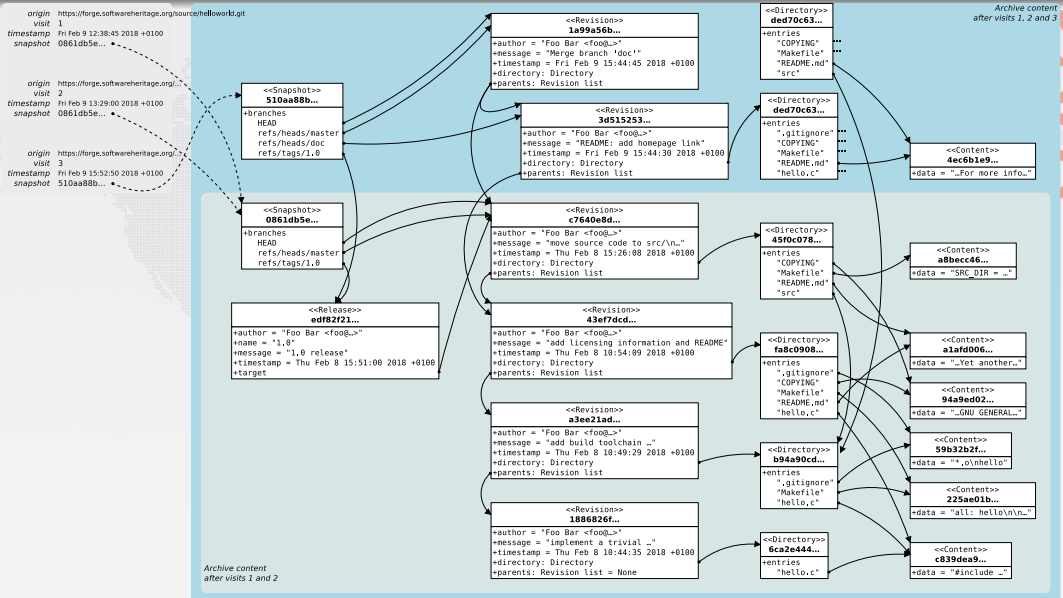
Organization

- non-profit
- multi-stakeholder

Data flow


Forges
GitHub lister
GitLab lister
Distros
Debian lister
PyPI lister
Package repos
CPAN
...
software origins
git
git
git
git
svn
svn
svn
hg
hg
hg
deb
deb
pypi
pypi
Git loader
Mercurial loader
Debian source package loader
PyPI loader
Software Heritage Archive
Merkle DAG
+
blob storage
Listing (full/incremental)
Scheduling
Loading & deduplication

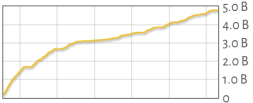
The archive: a (giant) Merkle DAG



Archive coverage — archive.softwareheritage.org

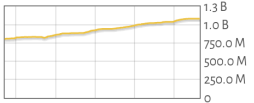
Source files

5,603,274,836



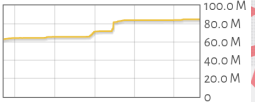
Commits

1,248,389,319



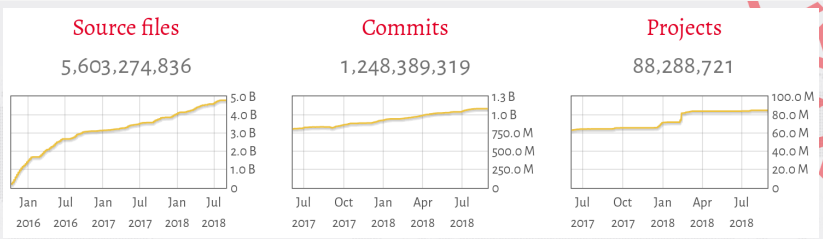
Projects

88,288,721



GitHub · debian · GitLab · Google code · {P} · python Package Index · GITORIOUS · GNU · HAL archives-ouvertes.fr · Inria inventeurs du monde numérique

Archive coverage — archive.softwareheritage.org



GitHub · debian · GitLab · Google code · {P} · python Package Index · GITORIOUS · GNU · HAL archives-ouvertes.fr · Inria inventeurs du monde numérique

- 200 TB (compressed) blobs, 6 TB database (as a graph: 10 B nodes + 100 B edges)
- The *richest* public source code archive, … and growing daily!

Archive all public code

How to maximize the impact of public code

1. Public money? Public code! (i.e., FOSS)
2. Go where the developers are
   - software distribution
   - community building
3. Archive for the long term
   - software developer platforms *are not* archives

Archive all public code

How to maximize the impact of public code

1. Public money? Public code! (i.e., FOSS)
2. Go where the developers are
   - software distribution
   - community building
3. Archive for the long term
   - software developer platforms *are not* archives

Reach out in both space (community) and time (posterity).

Find all public code

Let's build the global library of public code

- Make sure public code is properly tagged
  - general software ontologies, and/or
  - PA-specific ones (e.g., publiccode.yml)
- Index metadata from archived public code
- Make it searchable!

Find all public code

Let's build the global library of public code

- Make sure public code is properly tagged
  - general software ontologies, and/or
  - PA-specific ones (e.g., publiccode.yml)
- Index metadata from archived public code
- Make it searchable!

Work in progress...

... but already working on Python, JavaScript, Ruby, and Java metadata

`https://www.softwareheritage.org/2019/05/28/mining-software-metadata-for-80-m-projects-and-even-more/`

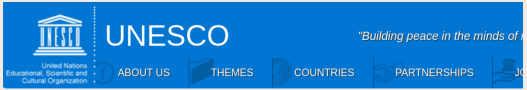
Raising Awareness

April 3rd 2017, Unesco Inria agreement

Inria
INVENTEURS DU MONDE NUMÉRIQUE

UNESCO

November 2018, Unesco Inria expert call



Home › All News › Experts call for greater recognition of software source code as heritage for sustainable development

Experts call for greater recognition of software source code as heritage for sustainable development

16 November 2018

Growing Support

Sharing the vision

Donors, members, sponsors

Inria
INVENTEURS DU MONDE NUMÉRIQUE

>= 100Ke/year	Microsoft, intel, SOCIETE GENERALE
>= 50Ke/year	CAST, Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE, Google
>= 25Ke/year	DANS, NOKIA Bell Labs
>= 10Ke/year	GitHub, UQÀM, FOSSID

Research collaboration

Qwant

source code search engine

Global network

FOSSID

- first independent mirror
- increased reliability

Software Heritage: the largest archive of reusable FOSS

Software Heritage

www.softwareheritage.org @swheritage

Archive all public code

- release public code
- archive it for posterity

Find all public code

- tag public code
- find it on Software Heritage

Come in, we're open

https://www.softwareheritage.org/community/developers/