Software Heritage

Archive All the Source Code for Better Compliance

Stefano Zacchiroli
Software Heritage — zack@upsilon.cc, @zacchiro

7 December 2018
Open Compliance Summit
Yokohama, Japan

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Talk spoilers

- Software Heritage: the ultimate source code archive
- 3 highlights on how Software Heritage relates to Open Compliance
  1. complete corresponding source code deposit
  2. persistent identifiers for source code artifacts
  3. a mutualized, open data, FOSS provenance database

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Our mission

Collect, preserve and share the *source code* of *all the software* that is publicly available.

Past, present and future

Preserving the past, *enhancing* the present, *preparing* the future.

Our principles

Cultural Heritage | **Industry** | **Research** | **Education**




Software Heritage

Open approach

- open source
- transparency

In for the long haul

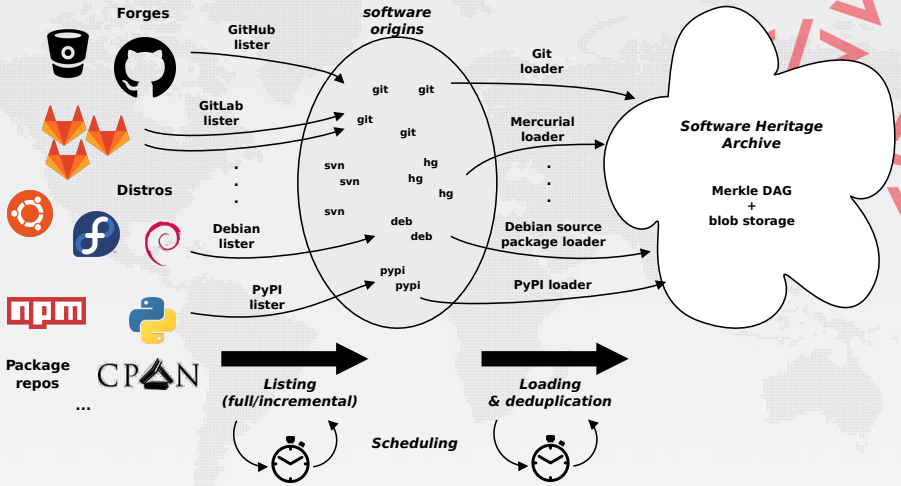
- non profit
- replication & mirrors

Breaking news: the first *independent* mirror is coming

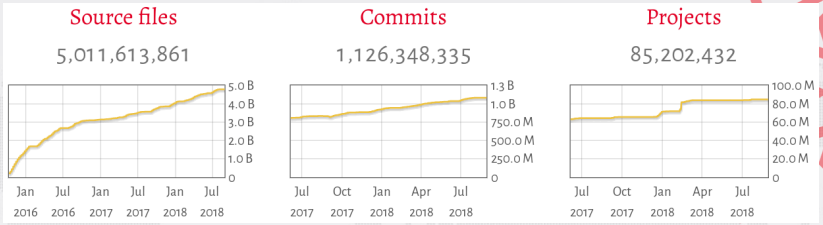
FOSSID Establishes First Independent Mirror of World's Largest Source Code Archive

by Fredrik Ehrenstrale | Dec 6, 2018 | Blog post | 0 comments

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

Archive coverage — archive.softwareheritage.org

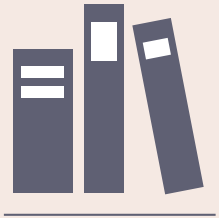


GitHub debian GitLab Google code {P} python™ Package Index GITORIOUS GNU HAL archives-ouvertes.fr Inria inventeurs du monde numérique

- 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!

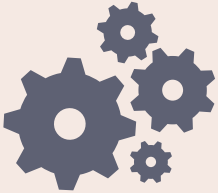
Accessing the archive

Browse



- `https://archive.softwareheritage.org/browse`
- way back machine for software source code

Web API



- `https://archive.softwareheritage.org/api`
- point-wise navigation of the archive as a graph

Case study: Complete Corresponding Source code distribution

- if you ship copylefted code, you also need to ship Complete Corresponding Source (CCS) code
- can you *outsource* hosting & distribution of CCS archives?
  yes (IANAL, etc.), e.g.,

> *GPL FAQ: Can I put the binaries on my Internet server and put the source on a different Internet site?*
>
> - *[v3] Yes. Section 6(d) allows this. However, you must provide clear instructions people can follow to obtain the source, and you must take care to make sure that the source remains available for as long as you distribute the object code.*
> - *[v2] The GPL says you must offer access to copy the source code "from the same place"; that is, next to the binaries. However, if you make arrangements with another site to keep the necessary source code available, and put a link or cross-reference to the source code next to the binaries, we think that qualifies as "from the same place".*

Open Compliance — Highlight #1

Software Heritage can permanently archive, host, and distribute CCS archives for you

Depositing source code to Software Heritage

Deposit service

- complement regular (pull) crawling of forges and distributions
- restricted access (i.e., not a warez dumpster!)
- `deposit.softwareheritage.org`

Tech bits

- SWORD 2.0 compliant server, for digital repositories interoperability
- RESTful API for deposit and monitoring, with CLI wrapper

Prepare a deposit

Prepare source code tarball

```
$ tar caf software.tar.gz /path/to/software/
```

Associate metadata

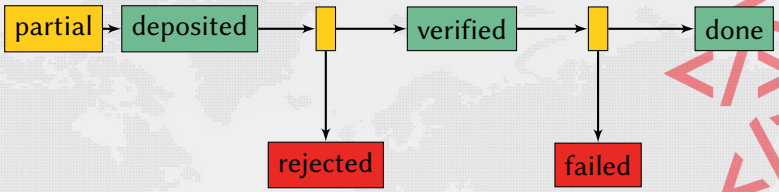
```
$ cat > software.tar.gz.metadata.xml
<?xml version="1.0"?>
<entry xmlns="http://www.w3.org/2005/Atom"
       xmlns:codemeta="https://doi.org/10.5063/SCHEMA/CODEMETA-2.0">
  <title>Je suis GPL</title>
 <codemeta:url>https://forge.softwareheritage.org/source/jesuisgpl/</codemeta:url>
  <codemeta:author>
    <codemeta:name>Stefano Zacchiroli</codemeta:name>
    <codemeta:jobTitle>Maintainer</codemeta:jobTitle>
  </codemeta:author>
</entry>
^D
```

Send a deposit

```
$ swh-deposit --username 'name' --password 'pass' \
      --archive software.tar.gz

{
  'deposit_id': '11',
  'deposit_status': 'deposited',
  'deposit_date': 'Jan. 30, 2018, 9:37 a.m.'
}
```

Ingestion status




```
$ swh-deposit --username 'name' --pass 'secret' \
      --deposit-id '11' --status

{
  'deposit_id': 11,
  'deposit_status': 'done',
  'deposit_status_detail': The deposit has been successfully loaded
   into the Software Heritage archive',
 'deposit_swh_id': 'swh:1:rev:a86747d201ab8f8657d145df4376676d5e47cf9f'
}
```

Access a deposit

After ingestion a deposit becomes an integral, permanent part of the archive.

- it has a persistent identifier
  - e.g., `swh:1:rev:a86747d201ab8f8657d145df4376676d5e47cf9f`
- it can be browsed online at `archive.softwareheritage.org`
  - e.g., `https://archive.softwareheritage.org/browse/swh:1:rev:a86747d201ab8f8657d145df4376676d5e47cf9f`

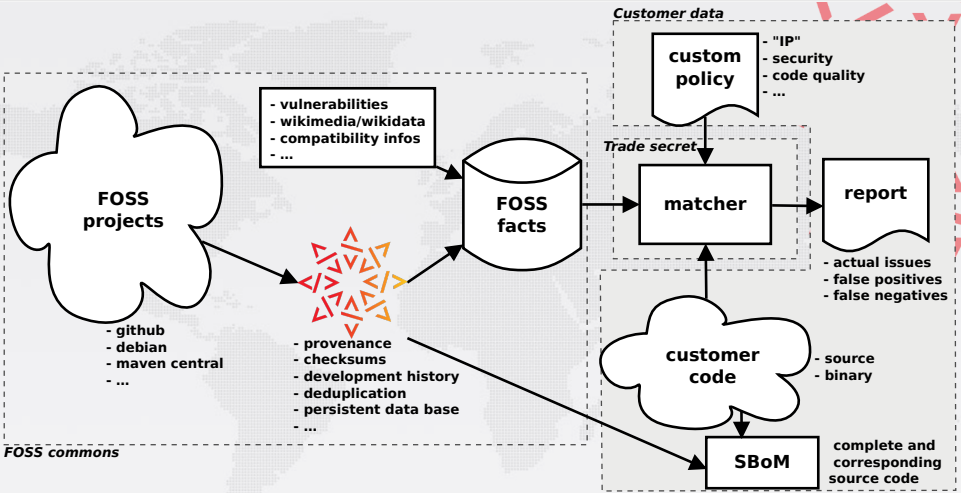
Software Heritage Persistent Identifiers (PIDs)

- Di Cosmo, Gruenpeter, Zacchiroli. *Identifiers for Digital Objects: the Case of Software Source Code Preservation*. In proceedings of iPRES 2018.
- `docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html`

Open Compliance — Highlight #2

Software Heritage assigns persistent identifiers to all archived source code artifacts

Towards an open provenance database



Open Compliance — Highlight #3

Software Heritage: a persistent, open data, provenance database for *all* FOSS

Wrapping up

Software Heritage…

- … collects, preserves, and shares the entire software commons
- … exhibits many synergies with the vision and goals of Open Provenance
- … is a non-profit endeavor, creating mutualized infrastructures and services

Come and join us

- `https://www.softwareheritage.org`
- `https://archive.softwareheritage.org`
- `https://www.softwareheritage.org/support/sponsors/`

… or just talk to me! :-) Stefano Zacchiroli / zack@upsilon.cc / @zacchiro