Software Heritage

Browsing 20 years of FOSS, and then some

Stefano Zacchiroli

Software Heritage — zack@upsilon.cc

18 July 2018
OSCON, Portland OR

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Free/Open Source Software is everywhere

Software source code is *special*

Harold Abelson, Structure and Interpretation of Computer Programs

"Programs must be written for people to read, and only incidentally for machines to execute."

Quake 2 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;
}
```

Net. queue in Linux (excerpt)

```
/*
 * SFB uses two B[l][n] : L x N arrays of bins (L levels, N bins per level)
 * This implementation uses L = 8 and N = 16
 * This permits us to split one 32bit hash (provided per packet by rxhash or
 * external classifier) into 8 subhashes of 4 bits.
 */
#define SFB_BUCKET_SHIFT 4
#define SFB_NUMBUCKETS  (1 << SFB_BUCKET_SHIFT) /* N bins per Level */
#define SFB_BUCKET_MASK (SFB_NUMBUCKETS - 1)
#define SFB_LEVELS      (32 / SFB_BUCKET_SHIFT) /* L */

/* SFB algo uses a virtual queue, named "bin" */
struct sfb_bucket {
        u16             qlen; /* length of virtual queue */
        u16             p_mark; /* marking probability */
};
```

Len Shustek, Computer History Museum

"Source code provides a view into the mind of the designer."

Our Software Commons

Definition (Commons)

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. https://en.wikipedia.org/wiki/Commons

Definition (Software Commons)

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.* [...]

https://en.wikipedia.org/wiki/Software_Commons

Our Software Commons

Definition (Commons)

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. https://en.wikipedia.org/wiki/Commons

Definition (Software Commons)

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.* [...]

https://en.wikipedia.org/wiki/Software_Commons

Source code is *a precious part* of our commons

are we taking care of it?

Software Heritage in a nutshell

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

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

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* that is publicly available

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

Reference catalog

find and reference all 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* that is publicly available

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

Reference catalog

find and reference all source code

Universal archive

preserve the software commons

Software Heritage in a nutshell

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

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

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

Reference catalog

find and reference all source code

Universal archive

preserve the software commons

Research infrastructure

enable analysis of all public software

Core principles

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




Software Heritage

Core principles

Open approach

- 100% Free Software
- transparency

In for the long haul

- replication
- non profit

Archiving goals

Targets: VCS repositories & source code releases (e.g., tarballs)

We DO archive

- file content (= blobs)
- revisions (= commits), with full metadata
- releases (= tags), ditto
- where (origin) & when (visit) we found any of the above

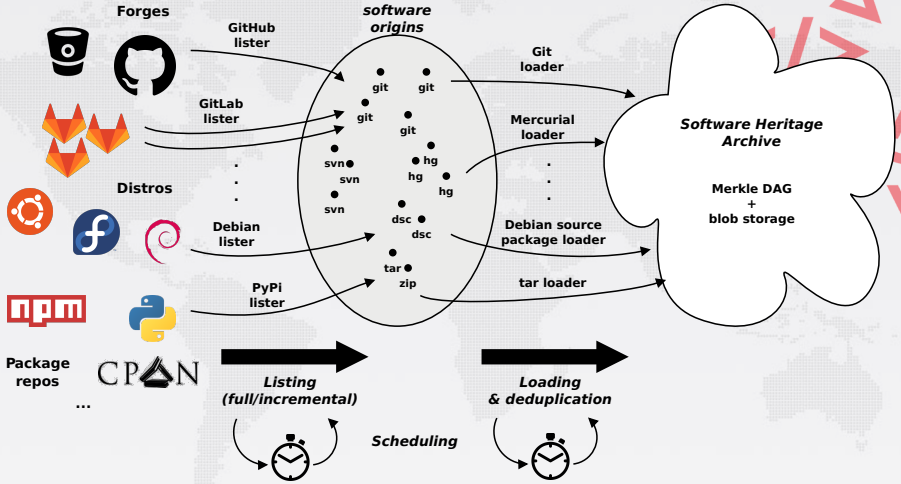
… in a VCS-/archive-agnostic canonical data model

We DON'T archive

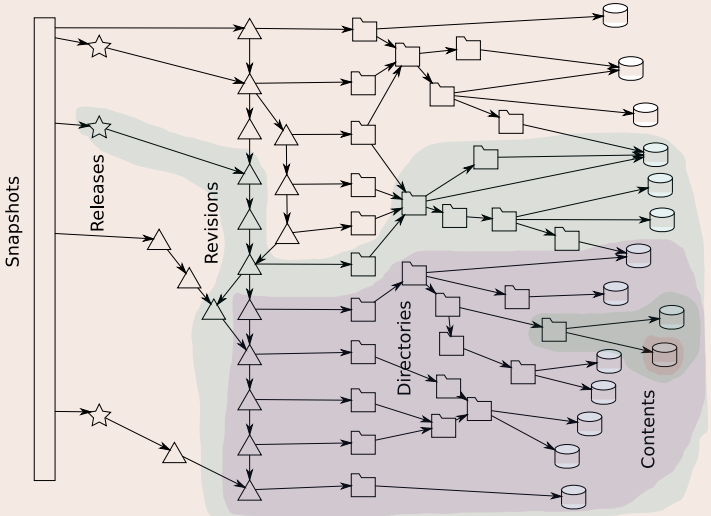
- homepages, wikis
- BTS/issues/code reviews/etc.
- mailing lists

Long term vision: play our part in a *"semantic wikipedia of software"*

Data flow


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

The archive: a (giant) Merkle DAG

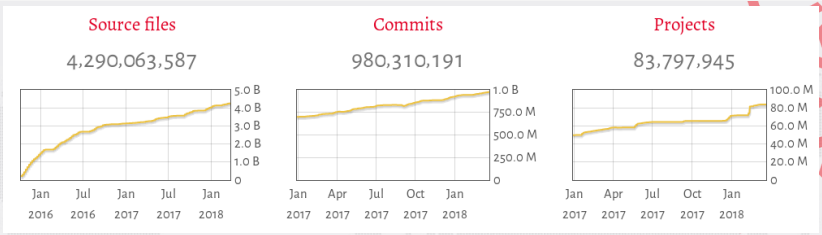


Archive coverage

Current sources

- live: GitHub, Debian
- one-off: Gitorious, Google Code, GNU
- WIP: GitLab, PyPI, Bitbucket

Archive coverage

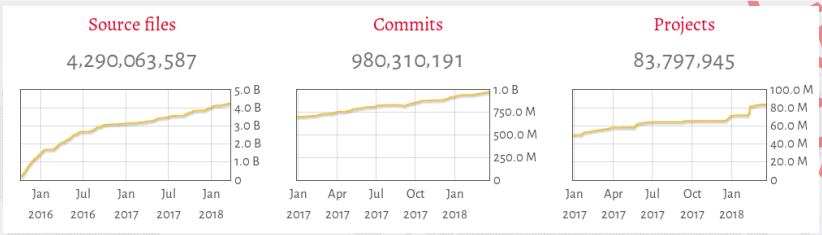


Current sources

- live: GitHub, Debian
- one-off: Gitorious, Google Code, GNU
- WIP: GitLab, PyPI, Bitbucket

175 TB (compressed) blobs, 6 TB database (as a graph: 10 B nodes + 100 B edges)

Archive coverage



Current sources

- live: GitHub, Debian
- one-off: Gitorious, Google Code, GNU
- WIP: GitLab, PyPI, Bitbucket

175 TB (compressed) blobs, 6 TB database (as a graph: 10 B nodes + 100 B edges)

The *richest* public source code archive, … and growing daily!

Web API

RESTful API to programmatically access the Software Heritage archive
`https://archive.softwareheritage.org/api/`

Features

- pointwise browsing of the archive
  - … snapshots ⟶ revisions ⟶ directories ⟶ contents …
- full access to the metadata of archived objects
- crawling information
  - *when have you last visited this Git repository I care about?*
  - *where were its branches/tags pointing to at the time?*

Endpoint index

`https://archive.softwareheritage.org/api/1/`

Bulk download

Vault service

- source code is thoroughly deduplicated within the Software Heritage archive
- bulk download of large artefacts (e.g., a Linux kernel release) requires collecting millions of objects
- the Software Heritage Vault cooks and caches source code bundles for bulk download needs

Tech bits

- RESTful API to request downloads, notifications, and monitoring
- `docs.softwareheritage.org/devel/swh-vault`

Web user interface

Browser-based interface to browse the Software Heritage archive
https://archive.softwareheritage.org/browse/

Features

- all REST API features, but good looking :-)
  - browsing: snapshots → revisions → directories → contents ...
  - access to metadata and crawling information
- origin search, as full text indexing of origin URLs
- bulk download, via integration with the Vault

You can help!

Coding

★★	Web UI improvements
★	loaders/listers for unsupported VCS/forges
★★★	developer documentation

`https://docs.softwareheritage.org/devel/`

You can help!

Coding

★★	Web UI improvements
★	loaders/listers for unsupported VCS/forges
★★★	developer documentation

`https://docs.softwareheritage.org/devel/`

Community

★★★	spread the world, help us with sustainability
★★	document endangered source code

`wiki.softwareheritage.org/index.php?title=Suggestion_box`

You can help!

Coding

★★	Web UI improvements
★	loaders/listers for unsupported VCS/forges
★★★	developer documentation

`https://docs.softwareheritage.org/devel/`

Community

★★★	spread the world, help us with sustainability
★★	document endangered source code

`wiki.softwareheritage.org/index.php?title=Suggestion_box`

Join us

- `www.softwareheritage.org/jobs` — job openings
- `wiki.softwareheritage.org/index.php?title=Internship` — internships

Conclusion

- we have come a long way
- we are now at a turning point in the history of FOSS
- we should preserve our technical legacy as solid foundation for the future
- Software Heritage, as a reference archive of all FOSS ever written, is here to help

Come in, we're open!

`www.softwareheritage.org`
`forge.softwareheritage.org`