

Software Heritage

Preserving the Free Software Commons

Roberto Di Cosmo and Stefano Zacchiroli

roberto@dicosmo.org, zack@upsilon.cc

4 February 2017

FOSDEM'17

Brussels, Belgium



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Free 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 also 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.”

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 is spread all around



Fashion victims

- many disparate development platforms
- a myriad places where distribution may happen
- projects tend to migrate from one place to another over time

Software is spread all around



Fashion victims

- many disparate development platforms
- a myriad places where distribution may happen
- projects tend to migrate from one place to another over time

Where is the place ...

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

A word cloud centered on a world map background, with various terms related to software fragility. The most prominent words are 'damage', 'disaster', 'malicious', 'deletion', 'obsolete', and 'attack'. Other visible words include 'dependencies', 'format', 'corruption', 'dangling', 'wear', 'encryption', 'reference', 'storage', 'media', 'aging', and 'tear'. The words are in various colors and orientations, with some appearing vertically or at an angle.

damage
disaster
malicious
deletion
obsolete
dependencies
attack
format
corruption
dangling
wear
encryption
reference
storage
media
aging
tear

Like all digital information, FOSS is fragile

- inconsiderate and/or malicious code loss (e.g., Code Spaces)
- business-driven code loss (e.g., Gitorious, Google Code)
- for obsolete code: physical media decay (data rot)



A word cloud of terms related to software fragility and digital information loss. The most prominent words are 'damage', 'disaster', 'malicious', 'obsolete', 'deletion', and 'attack'. Other visible words include 'dependencies', 'aging', 'tear', 'media', 'reference', 'storage', 'dangling', 'wear', 'corruption', 'encryption', and 'format'. The words are arranged in a circular pattern, with some overlapping. The background features a faint world map and several large, stylized arrows pointing outwards.

Like all digital information, FOSS is fragile

- inconsiderate and/or malicious code loss (e.g., Code Spaces)
- business-driven code loss (e.g., Gitorious, Google Code)
- for obsolete code: physical media decay (data rot)

Where is the archive...

where we go if (a repository on) GitHub or GitLab goes away?

Software is missing its own Research Infrastructure



Photo: ALMA(ESO/NAOJ/NRAO), R. Hills

A wealth of software research on crucial issues...

- safety, security; test, verification, proof;
- software engineering, software evolution;
- empirical and big data studies;

Software is missing its own Research Infrastructure



Photo: ALMA(ESO/NAOJ/NRAO), R. Hills

A wealth of software research on crucial issues...

- safety, security; test, verification, proof;
- software engineering, software evolution;
- empirical and big data studies;

If you study the stars, you go to Atacama...

... where is the *very large telescope* of source code?



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Our mission

Collect, **preserve** and **share** the *source code* of *all the software* that lies at the heart of our culture and our society.

Past, present and future

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

Cultural Heritage



Industry



Research



Education



Software Heritage

Cultural Heritage



Industry



Research



Education



Software Heritage

Open approach

- 100% FOSS
- transparency

In for the long haul

- replication
- non profit

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

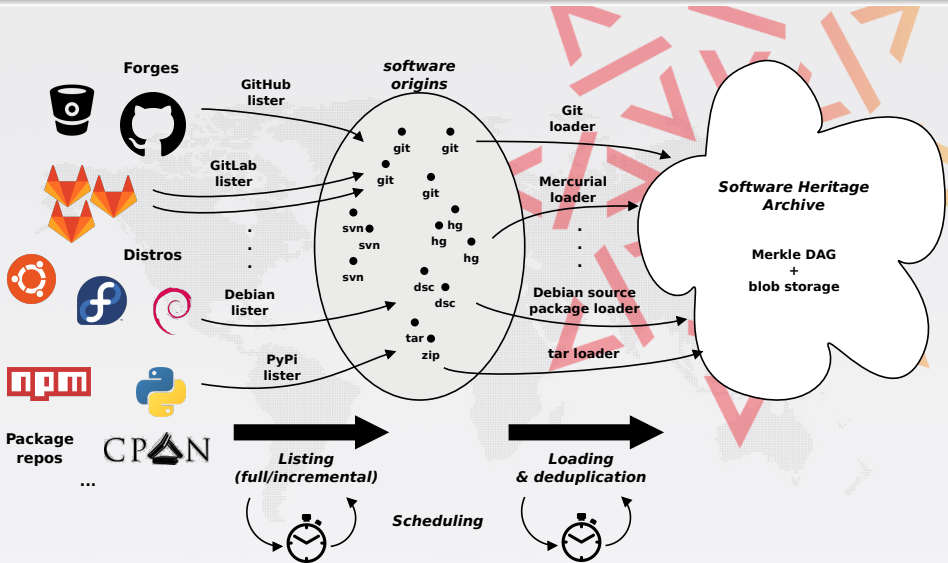
... in a VCS-/archive-agnostic **canonical data model**

We DON'T archive

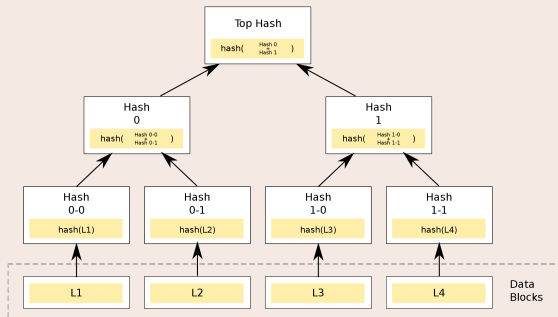
- homepages, wikis
- BTS/issues/code reviews/etc.
- mailing lists

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

Data flow



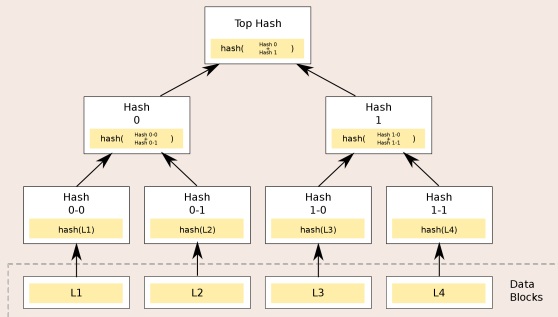
Merkle tree (R. C. Merkle, Crypto 1979)



Combination of

- tree
- hash function

Merkle tree (R. C. Merkle, Crypto 1979)




Combination of

- tree
- hash function

Classical cryptographic construction

- fast, parallel signature of large data structures
- widely used (e.g., Git, Bitcoin, IPFS, ...)
- built-in deduplication

Revisions

Details	Changes	Files
SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6		
Author: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)		
Committer: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)		
Subject: provenance.tasks: add the revision -> origin cache task		
Parent: fc3a8b59ca1df424d860f2c29ab07fee4dc35d10 : test_storage: properly pipeline origin and cont...		
provenance.tasks: add the revision -> origin cache task		
swh/storage/provenance/tasks.py  77		

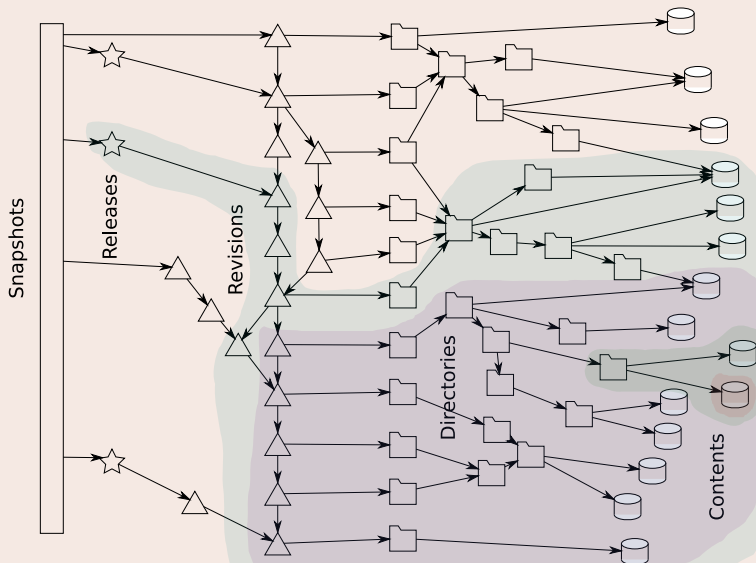
tree 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
parent fc3a8b59ca1df424d860f2c29ab07fee4dc35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

provenance.tasks: add the revision -> origin cache task

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6

Note: most object kinds currently have Git-compatible identifiers

The archive: a (giant) Merkle DAG



Our sources

- GitHub — full, up-to-date mirror
- Debian — daily snapshots of all suites since 2005–2015
- GNU — all releases as of August 2015
- Gitorious, Google Code — local copy (Archive Team & Google)

Our sources

- GitHub — full, up-to-date mirror
- Debian — daily snapshots of all suites since 2005–2015
- GNU — all releases as of August 2015
- Gitorious, Google Code — local copy (Archive Team & Google)

Some numbers



150 TB blobs, 6 TB database (as a graph: 5 B nodes + 50 B edges)

Our sources

- GitHub — full, up-to-date mirror
- Debian — daily snapshots of all suites since 2005–2015
- GNU — all releases as of August 2015
- Gitorious, Google Code — local copy (Archive Team & Google)

Some numbers



150 TB blobs, 6 TB database (as a graph: 5 B nodes + 50 B edges)

The *richest* source code archive already, ... and growing daily!

Web API (FOSDEM'17 release)

Fresh from the oven: first public version of our Web API

<https://archive.softwareheritage.org/api/>

Fresh from the oven: first public version of our Web API

<https://archive.softwareheritage.org/api/>

Features

- pointwise **browsing** of the Software Heritage archive
 - ... releases → 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?*

Fresh from the oven: first public version of our Web API

<https://archive.softwareheritage.org/api/>

Features

- pointwise **browsing** of the Software Heritage archive
 - ... releases → 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?*

Complete endpoint index

<https://archive.softwareheritage.org/api/1/>

A tour of the Web API — origins & visits

```
GET https://archive.softwareheritage.org/api/1/origin/ \
    git/url/https://github.com/hylang/hy
{ "id": 1,
  "origin_visits_url": "/api/1/origin/1/visits/",
  "type": "git",
  "url": "https://github.com/hylang/hy"
}
```

```
GET https://archive.softwareheritage.org/api/1/origin/ \
    1/visits/
[ ...,
  { "date": 1473851066.769266,
    "origin": 1,
    "origin_visit_url": "/api/1/origin/1/visit/13/",
    "status": "full",
    "visit": 13
  }, ...
]
```

A tour of the Web API — snapshots

```
GET https://archive.softwareheritage.org/api/1/origin/ \
    1/visit/13/
{ ...,
  "occurrences": { ...,
    "refs/heads/master": {
      "target": "b94211251...",
      "target_type": "revision",
      "target_url": "/api/1/revision/b94211251.../"
    },
    "refs/tags/0.10.0": {
      "target": "7045404f3...",
      "target_type": "release",
      "target_url": "/api/1/release/7045404f3.../"
    },
    }, ...
  },
  "origin": 1,
  "origin_url": "/api/1/origin/1/",
  "status": "full",
  "visit": 13
}
```

A tour of the Web API — revisions

```
GET https://archive.softwareheritage.org/api/1/revision/ \
6072557b6c10cd9a21145781e26ad1f978ed14b9/
{
  "author": {
    "email": "tag@pault.ag",
    "fullname": "Paul Tagliamonte <tag@pault.ag>",
    "id": 96,
    "name": "Paul Tagliamonte"
  },
  "committer": { ... },
  "date": "2014-04-10T23:01:11-04:00",
  "committer_date": "2014-04-10T23:01:11-04:00",
  "directory": "2df4cd84e...",
  "directory_url": "/api/1/directory/2df4cd84e.../",
  "history_url": "/api/1/revision/6072557b6.../log/",
  "merge": false,
  "message": "0.10: The Oh f*ck it's PyCon release",
  "parent_urls": [ "/api/1/revision/10149f66e.../" ],
  "parents": [ "10149f66e..." ],
  ...
}
```

A tour of the Web API — contents

```
GET https://archive.softwareheritage.org/api/1/content/ \
    adc83b19e793491b1c6ea0fd8b46cd9f32e592fc/
{
  "data_url": "/api/1/content/sha1:adc83b19e.../raw/",
  "filetype_url": "/api/1/content/sha1:.../filetype/",
  "language_url": "/api/1/content/sha1:.../language/",
  "length": 1,
  "license_url": "/api/1/content/sha1:.../license/",
  "sha1": "adc83b19e...",
  "sha1_git": "8b1378917...",
  "sha256": "01ba4719c...",
  "status": "visible"
}
```

A tour of the Web API — contents

```
GET https://archive.softwareheritage.org/api/1/content/ \
    adc83b19e793491b1c6ea0fd8b46cd9f32e592fc/
{
  "data_url": "/api/1/content/sha1:adc83b19e.../raw/",
  "filetype_url": "/api/1/content/sha1:.../filetype/",
  "language_url": "/api/1/content/sha1:.../language/",
  "length": 1,
  "license_url": "/api/1/content/sha1:.../license/",
  "sha1": "adc83b19e...",
  "sha1_git": "8b1378917...",
  "sha256": "01ba4719c...",
  "status": "visible"
}
```

- rate limits apply throughout the API
- blob download not available yet

Features...

- (done) **lookup** by content hash
- **browsing**: "wayback machine" for archived code
 - (done) via Web API
 - (todo) via Web UI
- (todo) **download**: `wget / git clone` from the archive
- (todo) **provenance information** for all archived content
- (todo) **full-text search** on all archived source code files

Features...

- (done) **lookup** by content hash
- **browsing**: "wayback machine" for archived code
 - (done) via Web API
 - (todo) via Web UI
- (todo) **download**: `wget / git clone` from the archive
- (todo) **provenance information** for all archived content
- (todo) **full-text search** on all archived source code files

... and much more than one could possibly imagine

all the world's software development history in a single graph!

You can help!

Coding

- `www.softwareheritage.org/community/developers/`
- `forge.softwareheritage.org` – our own code

Current development priorities

- ★★★ listers for unsupported forges, distros, pkg. managers
- ★★★ loaders for unsupported VCS, source package formats
- ★★ Web UI: eye candy wrapper around the Web API
- ★ content indexing and search

... *all* contributions equally welcome!

You can help!

Coding

- www.softwareheritage.org/community/developers/
- forge.softwareheritage.org – our own code

Current development priorities

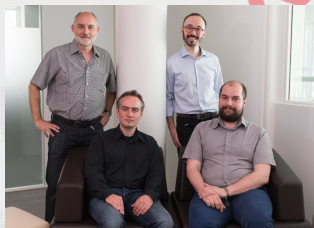
- ★★★ listers for unsupported forges, distros, pkg. managers
- ★★★ loaders for unsupported VCS, source package formats
- ★★ Web UI: eye candy wrapper around the Web API
- ★ content indexing and search

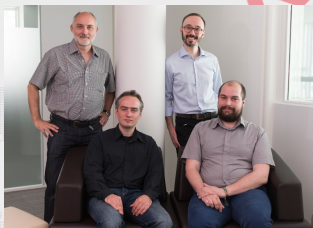
... *all* contributions equally welcome!

Join us

- www.softwareheritage.org/jobs – job openings
- wiki.softwareheritage.org – internships

The Software Heritage community

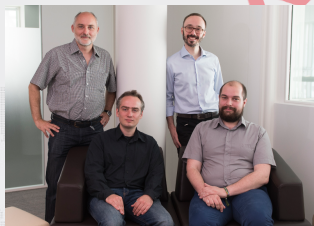




Inria as initiator



- .fr national CS research institution
- strong FOSS culture, W3C founding partner



Inria as initiator



- .fr national CS research institution
- strong FOSS culture, W3C founding partner

Early Partners and Supporters

Société Générale, Microsoft, Huawei, Nokia Bell Labs, DANS, ACM, Adullact, Creative Commons, Eclipse, Free Software Foundation, Gandi, Open Source Initiative, GitHub, IEEE, OIN, OW2, Software Freedom Conservancy, SFLC, The Document Foundation, The Linux Foundation, ...

Software Heritage is

- a *reference archive* of all FOSS ever written
- a unique *complement* for *development platforms*
- an international, open, nonprofit, *mutualized infrastructure*
- at the service of our community, at the service of society

Come in, we're open!

`www.softwareheritage.org` – *sponsoring, job openings*
`wiki.softwareheritage.org` – *internships, leads*
`forge.softwareheritage.org` – *our own code*

Questions?

Q: how about SHA1 collisions?

```
create domain sha1 as bytea
  check (length(value) = 20);
create domain sha1_git as bytea
  check (length(value) = 20);
create domain sha256 as bytea
  check (length(value) = 32);

create table content (
  sha1          sha1 primary key,
  sha1_git      sha1_git not null,
  sha256        sha256 not null,
  length        bigint not null,
  ctime         timestamptz not null default now(),
  status        content_status not null default 'visible',
  object_id     bigserial
);

create unique index on content(sha1_git);
create unique index on content(sha256);
```


Q: do you archive *only* Free Software?

- We only crawl origins *meant* to host source code (e.g., forges)
- Most (~90%) of what we *actually* retrieve is textual content
- Our goal: archive *the entire Free Software commons*
- Large parts of what we retrieve is *already* Free Software, today
- Most of the rest *will become* Free Software in the long term
 - e.g., at copyright expiration