

Software Heritage

Archiving the Software Commons for Fun and Social Benefit

Stefano Zacchiroli

University Paris Diderot & Inria
zack@epsilon.cc

24 May 2017

Nexa Center for Internet & Society – Turin, Italy



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Software is everywhere



The source code matters!



"The source code for a work means the preferred form of the work for making modifications to it."

GPL Licence





“The source code for a work means the preferred form of the work for making modifications to it.”

GPL Licence

Hello World

The source code matters!



“The source code for a work means the preferred form of the work for making modifications to it.”

GPL Licence

Hello World

Program (excerpt of binary)

```
4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
4004f9: 90
4004fa: 5d
4004fb: c3
```



"The source code for a work means the preferred form of the work for making modifications to it."

GPL Licence

Hello World

Program (excerpt of binary)

```
4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
4004f9: 90
4004fa: 5d
4004fb: c3
```

Program (source code)

```
/* Hello World program */

#include<stdio.h>

void main()
{
    printf("Hello World");
}
```

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

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?



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 of terms related to software fragility and digital information loss. The words are arranged in a roughly circular pattern. The largest words are 'damage', 'disaster', 'malicious', 'obsolete', 'deletion', and 'format'. Other words include 'attack', 'dependencies', 'aging', 'media', 'tear', 'dangling', 'wear', 'corruption', 'encryption', 'reference', and 'storage'. The colors of the words vary, including shades of purple, blue, green, and orange.

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 'format'. Other visible words include 'attack', 'dependencies', 'aging', 'media', 'tear', 'dangling', 'wear', 'corruption', 'encryption', 'reference', and 'storage'. The words are arranged in a cluster, with some overlapping.



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.com goes away?



A wealth of software research on crucial issues...

- safety, security, test, verification, proof
- software engineering, software evolution
- big data, machine learning, empirical studies



A wealth of software research on crucial issues...

- safety, security, test, verification, proof
- software engineering, software evolution
- big data, machine learning, empirical studies

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

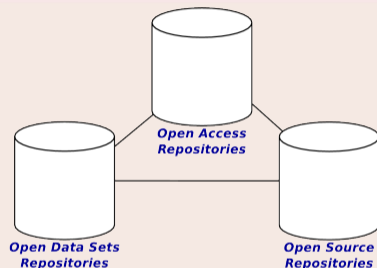
... where is the *very large telescope* of source code?

The (sad) state of software-related reproducibility

Reproducibility

- how are we doing, regarding reproducibility, in *software*?
e.g., Collberg et al., Measuring reproducibility in computer systems research (2014)
- not a CS-only issue, software is everywhere in science
- all round reproducibility is difficult; source code availability a requirement

Scientific knowledge preservation



Legenda (links are important!)

- articles: ArXiv, HAL, ...
- data: Zenodo, ...
- software: ?



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.



Software Heritage

Software Heritage

Collect

- discover
 - sources
- harvest
 - protocols
- ingest
 - VCS
 - data models

Software Heritage

Collect

- discover
 - sources
- harvest
 - protocols
- ingest
 - VCS
 - data models

Organise and Preserve

- enrich
 - metadata
- analyze
- replicate
 - locations
 - technologies
 - stakeholders

Software Heritage

Collect

- discover
 - sources
- harvest
 - protocols
- ingest
 - VCS
 - data models

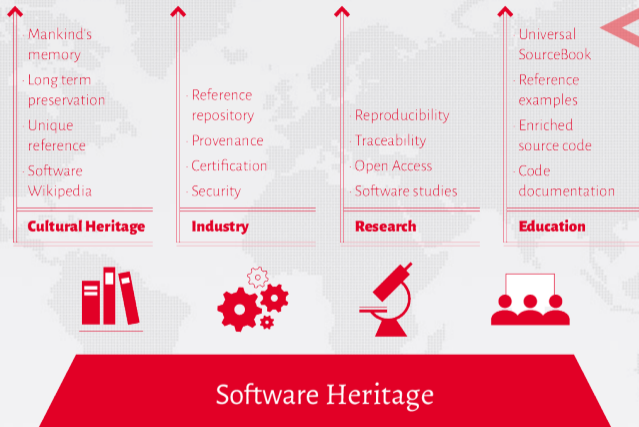
Organise and Preserve

- enrich
 - metadata
- analyze
- replicate
 - locations
 - technologies
 - stakeholders

Share

- download
- browse
- search
 - code
 - history
- watch
 - trends

one infrastructure to build them all



Our principles

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

The archive is ready and growing



Our current sources

- GitHub
- Debian, GNU
- WIP: Gitorious, Google Code, Bitbucket

The archive is ready and growing



Our current sources

- GitHub
- Debian, GNU
- WIP: Gitorious, Google Code, Bitbucket

The biggest source code archive already, ... and growing daily!

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!

Sharing the Software Heritage vision



See more

<http://www.softwareheritage.org/support/testimonials>

Inria
INVENTEURS DU MONDE NUMÉRIQUE



Microsoft



SOCIÉTÉ
GÉNÉRALE



HUAWEI

Data Archiving and Networked Services

DANS

NOKIA Bell Labs



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI INFORMATICA - SCIENZA E INGEGNERIA

April 3rd, 2017: landmark UNESCO/Inria agreement...

Inria
INVENTEURS DU MONDE NUMÉRIQUE



www.softwareheritage.org/?p=11623

Next step: 27-28 Sep 2017: UNESCO/Inria conference in Paris

You can help!

Coding

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

Working groups

- wiki.softwareheritage.org/index.php?title=Working_groups – **working groups**
 - ELIE (Ethical and Legal Issues and Environment) WG: legal and policy issues
 - copyright exceptions for archival, privacy issues in preserving source code and development metadata, etc.

Join us

- www.softwareheritage.org/jobs – **job openings**
- wiki.softwareheritage.org/index.php?title=Internships – **internships**

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, working groups*

`forge.softwareheritage.org` – *our own code*

Questions?

Archiving goals

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

We DO archive

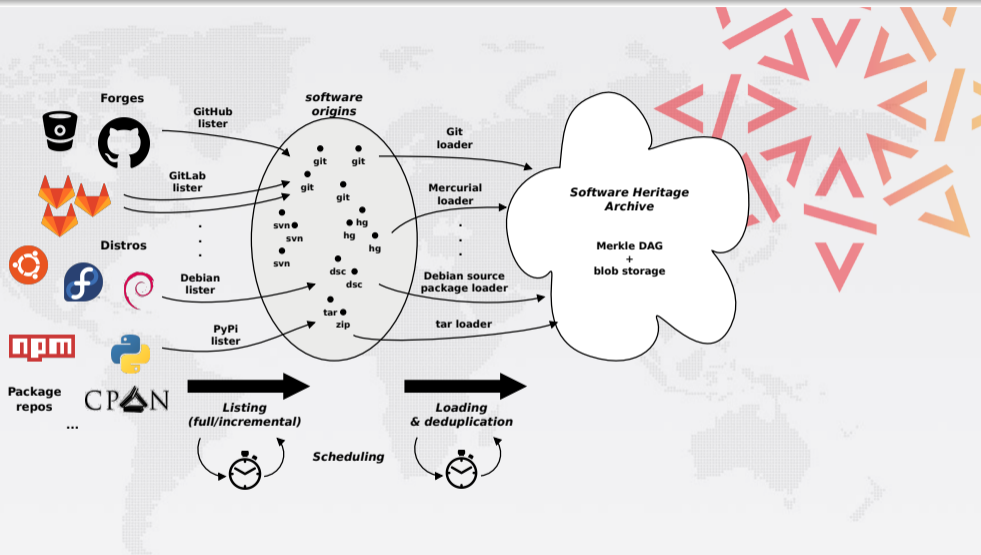
- file **content** (= blobs)
- **revisions** (= commits), with full metadata
- **releases** (= tags), ditto
- where (**origin**) & when (**visit**) we found any of the above

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

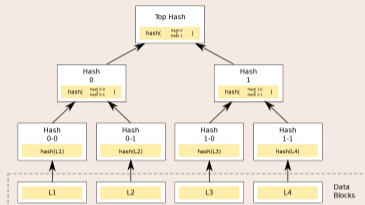
We DON'T archive

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

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



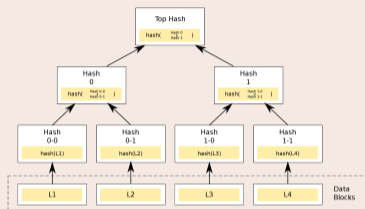
Merkle tree (R. C. Merkle, Crypto 1979)



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, blockchains, IPFS, ...)
- built-in deduplication

Revisions

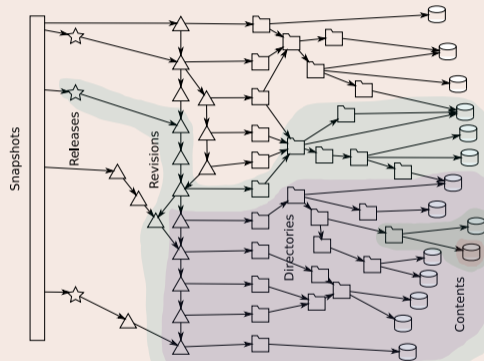
Details	Changes	Files
SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6		
Author: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)		
Committer: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)		
Subject: provenance.tasks: add the revision -> origin cache task		
Parent: 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](#)

The archive: a (giant) Merkle DAG



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": "2016-09-14T11:04:26.769266+00:00",
    "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",
  "parents": [ {
    "id": "10149f66e...",
    "url": "/api/1/revision/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"
}
```



```
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"
}
```

Caveats

- rate limits apply throughout the API
- blob download available for selected contents