

Software Heritage

The Great Library of Source Code

Stefano Zacchioli

Software Heritage – zack@upsilon.cc, [@zacchiro](https://twitter.com/zacchiro)

1 July 2019

Team per la Trasformazione Digitale – Roma, Italy



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 III 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.”



A word cloud of terms related to software fragility, including: damage, disaster, malicious, deletion, reference, storage, attack, obsolete, dependencies, dangling, wear, corruption, encryption, format, aging, media, and 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 words are arranged in a roughly circular pattern. The largest words are 'damage' (top), 'disaster' (left), 'malicious' (left), 'attack' (left), 'obsolete' (left), 'dependencies' (left), 'deletion' (right), 'reference' (right), 'storage' (right), 'format' (right), 'encryption' (right), 'corruption' (right), 'wear' (right), 'dangling' (right), 'aging' (left), 'media' (left), and 'tear' (left). The colors range from purple to green.

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 do we go if (a repository on) GitHub or GitLab.com goes away?



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 wealth of software research on crucial issues...

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

Software lacks its own research infrastructure



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?



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

Cultural Heritage



Industry



Research



Education



Software Heritage

Open approach

- open source
- transparency

In for the long haul

- non profit
- replication & mirrors

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

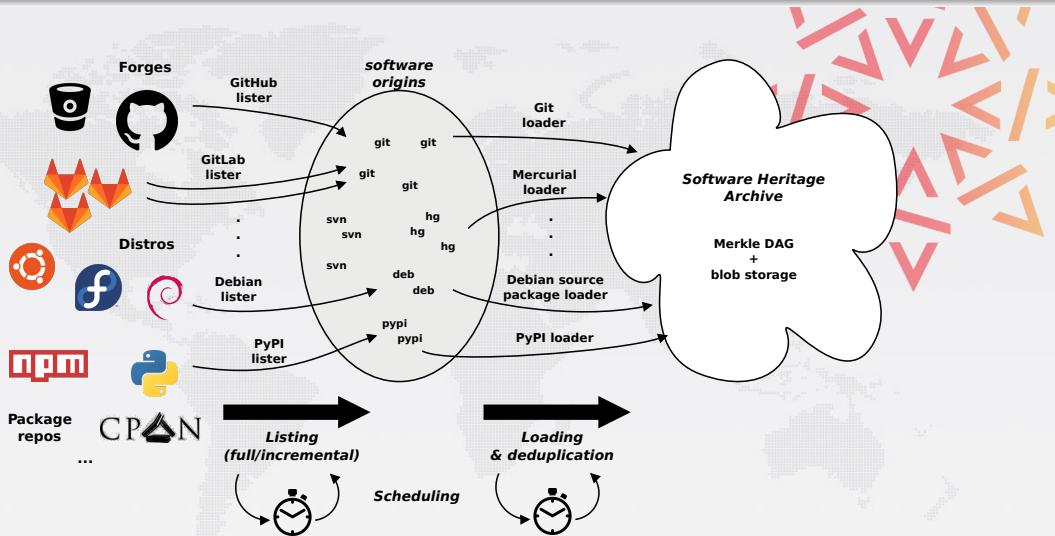
... in a VCS-/archive-agnostic **canonical data model**

We DON'T archive

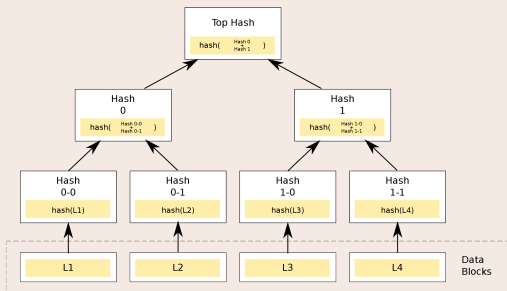
- homepages, wikis
- BTS/issues/code reviews/etc.
- mailing lists

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

Data flow



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

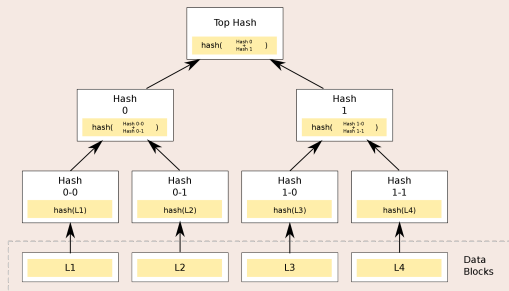


Combination of

- tree
- hash function

Merkle trees

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



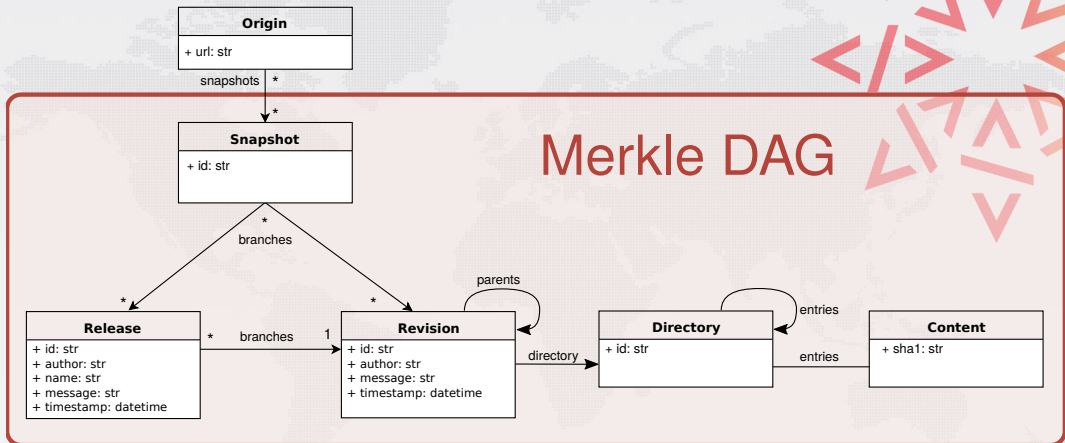
Combination of

- tree
- hash function

Classical cryptographic construction

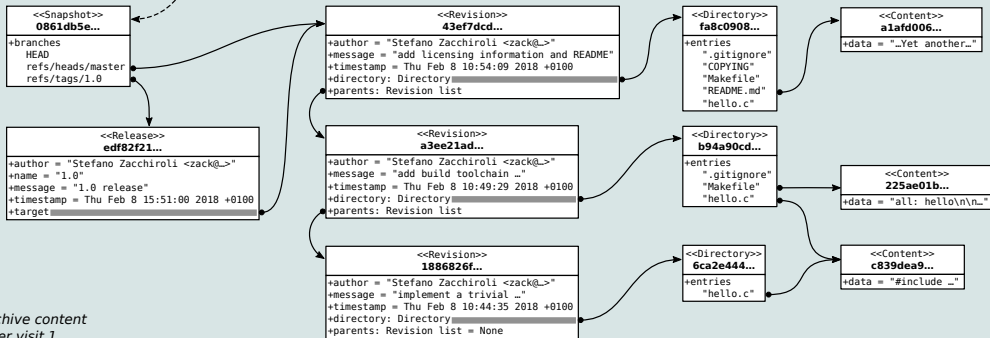
- fast, parallel signature of large data structures
- widely used (e.g., Git, blockchains, IPFS, ...)
- built-in deduplication

The archive: a (giant) Merkle DAG



The archive: a (giant) Merkle DAG

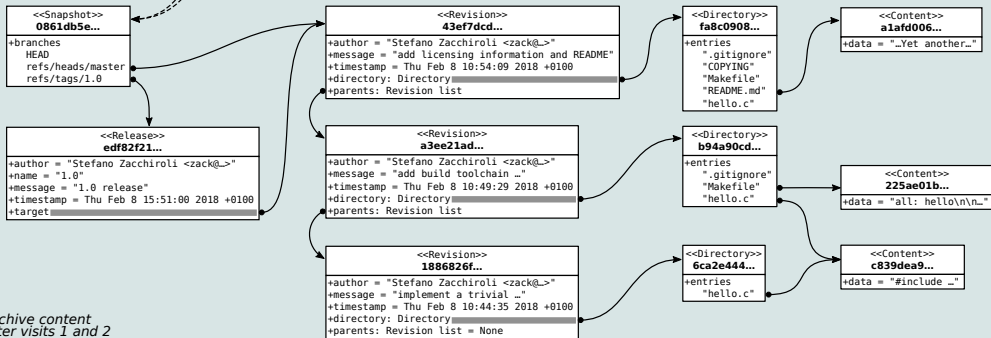
origin https://forge.softwareheritage.org/source/helloworld.git
visit 1
snapshot 0861db5e...
timestamp Fri Feb 9 12:38:45 2018 +0100



Archive content
after visit 1

The archive: a (giant) Merkle DAG

origin	visit	snapshot	timestamp
https://forge.softwareheritage.org/source/helloworld.git	1	0861db5e...	Fri Feb 9 12:38:45 2018 +0100
https://forge.softwareheritage.org/source/helloworld.git	2	0861db5e...	Fri Feb 9 13:29:00 2018 +0100



Archive content
after visits 1 and 2

The archive: a (giant) Merkle DAG

origin	visit	snapshot	timestamp
https://forge.softwareheritage.org/source/helloworld.git	1	0861db5e...	Fri Feb 9 12:38:45 2018 +0100
https://forge.softwareheritage.org/source/helloworld.git	2	0861db5e...	Fri Feb 9 13:29:00 2018 +0100
https://forge.softwareheritage.org/source/helloworld.git	3	510aa88b...	Fri Feb 9 15:52:50 2018 +0100

```
<<Snapshot>>
510aa88b...
+branches
HEAD
refs/heads/master
refs/heads/doc
...
refs/tags/1.0
```

```
<<Snapshot>>
0861db5e...
+branches
HEAD
refs/heads/master
refs/tags/1.0
```

```
<<Release>>
edf82f21...
+author = "Stefano Zacchioli <zack@...>"
+name = "1.0"
+message = "1.0 release"
+timestamp = Thu Feb 8 15:51:00 2018 +0100
+target
```

```
<<Revision>>
c7640e8d...
+author = "Stefano Zacchioli <zack@...>"
+message = "move source code to src/\n_"
+timestamp = Thu Feb 8 15:26:08 2018 +0100
+directory: Directory
+parents: Revision list
```

```
<<Revision>>
43ef7dcd...
+author = "Stefano Zacchioli <zack@...>"
+message = "add licensing information and README"
+timestamp = Thu Feb 8 10:54:09 2018 +0100
+directory: Directory
+parents: Revision list
```

```
<<Revision>>
a3ee21ad...
+author = "Stefano Zacchioli <zack@...>"
+message = "add build toolchain ..."
+timestamp = Thu Feb 8 10:49:29 2018 +0100
+directory: Directory
+parents: Revision list
```

```
<<Revision>>
1886826f...
+author = "Stefano Zacchioli <zack@...>"
+message = "implement a trivial ..."
+timestamp = Thu Feb 8 10:44:35 2018 +0100
+directory: Directory
+parents: Revision list = None
```

```
<<Directory>>
45f0c078...
+entries
"COPYING"
"Makefile"
"README.md"
"src"
```

```
<<Directory>>
fa8c0908...
+entries
".gitignore"
"COPYING"
"Makefile"
"README.md"
"hello.c"
```

```
<<Directory>>
b94a90cd...
+entries
".gitignore"
"Makefile"
"hello.c"
```

```
<<Directory>>
6ca2e444...
+entries
"hello.c"
```



Archive content
after visits 1, 2 and 3

```
<<Content>>
a1afd006...
+data = "...Yet another..."
```

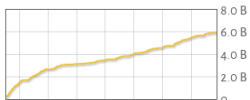
```
<<Content>>
225ae01b...
+data = "all: hello\n\n_"
```

```
<<Content>>
c839dea9...
+data = "#include _"
```

Archive content
after visits 1 and 2

Source files

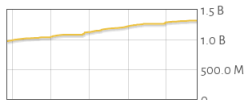
6,006,503,960



Jan Jul Jan Jul Jan Jul Jan
2016 2016 2017 2017 2018 2018 2019

Commits

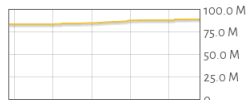
1,326,776,432



Apr Jul Oct Jan Apr
2018 2018 2018 2019 2019

Projects

89,301,694



Apr Jul Oct Jan Apr
2018 2018 2018 2019 2019

GitHub



GitLab



Google code



GITORIOUS



GNU

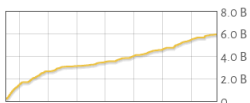
HAL
archives-ouvertes.fr

Inria
inventeurs du monde numérique



Source files

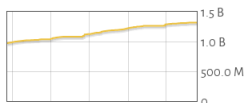
6,006,503,960



Jan Jul Jan Jul Jan Jul Jan
2016 2016 2017 2017 2018 2018 2019

Commits

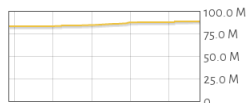
1,326,776,432



Apr Jul Oct Jan Apr
2018 2018 2018 2019 2019

Projects

89,301,694



Apr Jul Oct Jan Apr
2018 2018 2018 2019 2019

GitHub



GitLab



Google code



GITORIOUS



GNU

HAL
archives-ouvertes.fr

Inria
inventeurs du monde numérique



- ~400 TB (uncompressed) blobs, ~20 B nodes, ~280 B edges
- The *richest* public source code archive, ... and growing daily!

Browser-based interface to browse the Software Heritage archive

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

Features

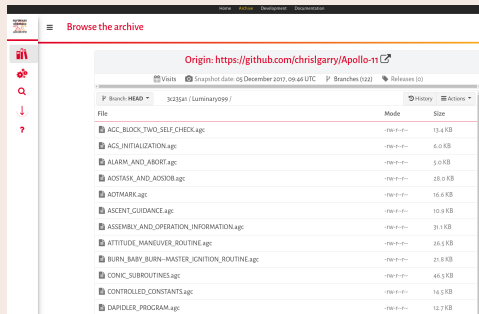
- all **REST API features**, but good looking :-)
 - browsing: snapshots → revisions → directories → contents ...
 - access to metadata and crawling information
- **origin search**, as full text indexing of origin URLs
- bulk **download**, via integration with the Vault

Example: the Apollo 11 source code

Margaret Hamilton



The Apollo 11 source code in SWH

A screenshot of the Apollo 11 source code repository in Software Heritage (SWH). The interface shows a file browser view for the repository. The origin is listed as <https://github.com/chrisgarry/Apollo-11>. The repository is identified as '3c235a1 / Luminary099 /'. A table lists the files in the repository, including their mode and size.

File	Mode	Size
AGC_BLOCK_TWO_SELF_CHECK.agc	-rw-r--r--	13.4 KB
ACS_INITIALIZATION.agc	-rw-r--r--	6.0 KB
ALARM_AND_ABORT.agc	-rw-r--r--	5.0 KB
AOSTASK_AND_AOSJOB.agc	-rw-r--r--	28.0 KB
AOTMARK.agc	-rw-r--r--	16.6 KB
ASCENT_GUIDANCE.agc	-rw-r--r--	10.9 KB
ASSEMBLY_AND_OPERATION_INFORMATION	-rw-r--r--	31.1 KB
ATTITUDE_MANEUVER_ROUTINE.agc	-rw-r--r--	26.5 KB
BURN_BABY_BURN-MASTER_IGNITION_ROUTINE.agc	-rw-r--r--	23.8 KB
CONIC_SUBROUTINES.agc	-rw-r--r--	46.5 KB
CONTROLLED_CONSTANTS.agc	-rw-r--r--	14.5 KB
DAPIDLER_PROGRAM.agc	-rw-r--r--	12.7 KB

Some pointers

- Entry point
- Burn, baby, burn!

Example: the Quake 3 source code

John Carmack



The Quake 3 source code in SWH

A screenshot of the Software Heritage (SWH) archive interface showing the Quake III Arena source code. The page title is "Browse the archive" and the origin is "https://github.com/id-Software/Quake-III-Arena". The page shows a list of files and directories, including "code", "common", "icc", "libs", "qasm", "q3map", "q3radiant", "ui", "COPYING.txt", and "README.txt". The "README.txt" file is highlighted in red, and the text "Quake III Arena GPL source release" is displayed below it.

File	Mode	Size
code	d----	
common	d----	
icc	d----	
libs	d----	
qasm	d----	
q3map	d----	
q3radiant	d----	
ui	d----	
COPYING.txt	-r--r--r--	14.8 KB
README.txt	-r--r--r--	8.8 KB

Some pointers

- Entry point
- What the f...

Software Heritage Graph dataset

Use case: large scale analyses of the most comprehensive corpus on the development history of free/open source software.

Dataset

- Relational representation of the full graph as a set of tables
- Available as open data: <https://doi.org/10.5281/zenodo.2583978>

Formats

- Local use: PostgreSQL dumps, or Apache Parquet files (~1 TiB each)
- Live usage: Amazon Athena (SQL-queriable)

References and sample queries



Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli

The Software Heritage Graph Dataset: Public software development under one roof

MSR 2019: Intl. Conf. on Mining Software Repositories, IEEE

non-paywalled preprint: <http://deb.li/swhmsr19>



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Contacts

- <https://www.softwareheritage.org>
- me: zack@upsilon.cc / [@zacchiro](#)



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Contacts

- <https://www.softwareheritage.org>
- me: zack@upsilon.cc / [@zacchiro](#)

Archive PA source code

- GitHub and forges are not archives !
- save code now
<https://tiny.cc/swh-save>



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Contacts

- <https://www.softwareheritage.org>
- me: zack@upsilon.cc / [@zacchiro](#)

Archive PA source code

- GitHub and forges are not archives !
- save code now
<https://tiny.cc/swh-save>

Mine PA code metadata

- [italia/publiccode.yml](#)
- project metadata mining
<https://tiny.cc/swh-mine>



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Contacts

- <https://www.softwareheritage.org>
- me: zack@upsilon.cc / [@zacchiro](#)

Archive PA source code

- GitHub and forges are not archives !
- save code now
<https://tiny.cc/swh-save>

Mine PA code metadata

- [italia/publiccode.yml](#)
- project metadata mining
<https://tiny.cc/swh-mine>

Track PA contriubs/impact

- how much FOSS is state-sponsored?
- what did PA code enable?

Most frequent first commit words

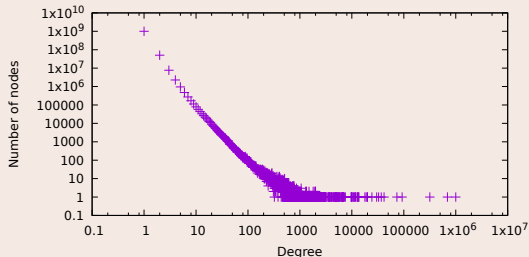
```
SELECT COUNT(*) AS c, word FROM (  
  SELECT LOWER(REGEXP_EXTRACT(FROM_UTF8(  
    message), '^\\w+')) AS word FROM revision)  
WHERE word != ''  
GROUP BY word ORDER BY COUNT(*) DESC LIMIT 5;
```

Count	Word
71'338'310	update
64'980'346	merge
56'854'372	add
44'971'954	added
33'222'056	fix

Fork arity

i.e., how often is a commit based upon?

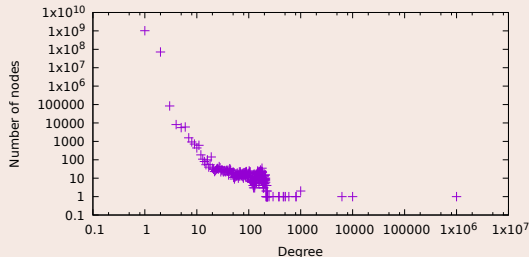
```
SELECT fork_deg, count(*) FROM (  
  SELECT id, count(*) AS fork_deg  
  FROM revision_history GROUP BY id) t  
GROUP BY fork_deg ORDER BY fork_deg;
```



Merge arity

i.e., how large are merges?

```
SELECT merge_deg, COUNT(*) FROM (  
  SELECT parent_id, COUNT(*) AS merge_deg  
  FROM revision_history GROUP BY parent_id) t  
GROUP BY deg ORDER BY deg;
```



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?*

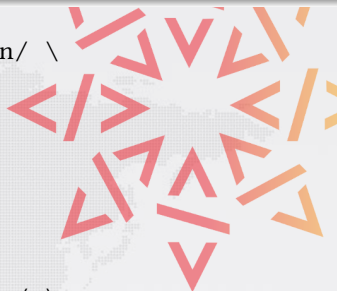
Endpoint index

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

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.../"  
  }  
]
```



```
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
- raw download available for textual contents

3rd party

- Debian, Puppet, Ceph
- PostgreSQL for metadata storage, with barman & pglogical
- Celery (RabbitMQ backend) for task scheduling
- Python3 and psycopg2 for the backend
- Django, Bootstrap, D3.js for Web stuff

in house

- *ad hoc* object storage (to avoid imposing tech to mirrors)
- data model implementation, listers, loaders, scheduler
- ~60 Git repositories (~20 Python packages, ~30 Puppet modules)
- ~60 kSLOC Python / ~12 kSLOC SQL / ~4 kSLOC Puppet
- licence choice: GPLv3 (backend) / AGPLv3 (frontend)

in house

- 2x hypervisors with ~20 VMs
- 1x high performance database server
- 2x dedicated storage server using
- 2x high density storage array (60 * 6TB => 300TB usable each)
- 3x nodes for a kafka+elasticsearch cluster

on Azure

- full object storage mirror
- full mirror of the database containing the graph
- workers for content indexing
- workers for download bundle preparation

classic FOSS development

- language: English
- development mailing list
<https://sympa.inria.fr/sympa/info/swh-devel>
- IRC
#swh-devel / FreeNode
- Forge
<https://forge.softwareheritage.org>
- Git, tasks, code review, etc.

for more information

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