Browsing the Free Software Commons

Stefano Zacchiroli

University Paris Diderot & Inria — zack@upsilon.cc

24 March 2018
LibrePlanet
Boston, MA, USA

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

1. The Free Software commons
2. Software Heritage
3. Accessing the archive
4. Getting involved

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 is fragile

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

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)

Software is fragile

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?

Outline

1. The Free Software commons
2. Software Heritage
3. Accessing the archive
4. Getting involved

The Software Heritage Project

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

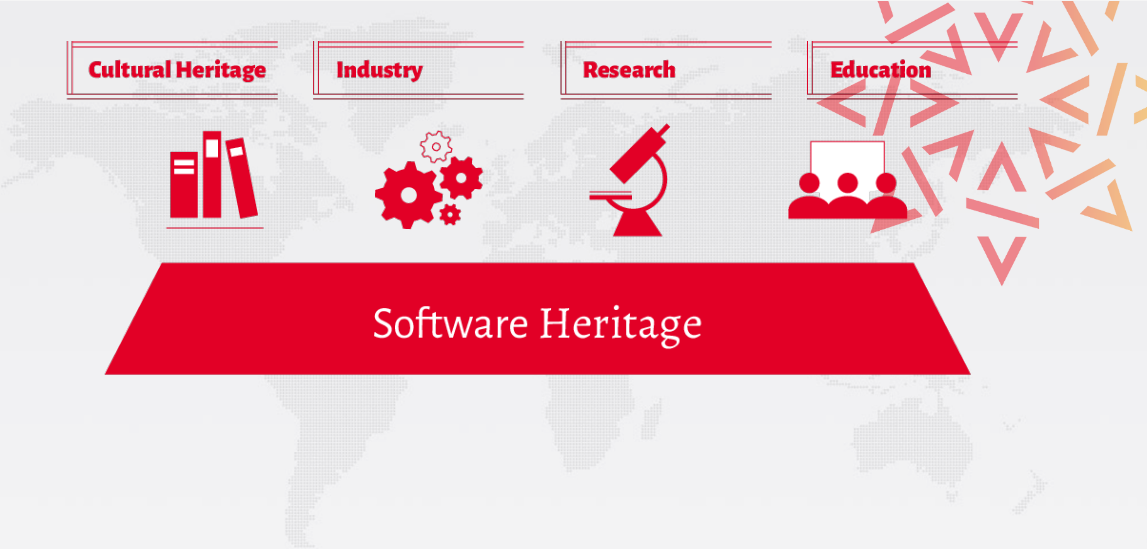
Our mission

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.

Core principles



Cultural Heritage

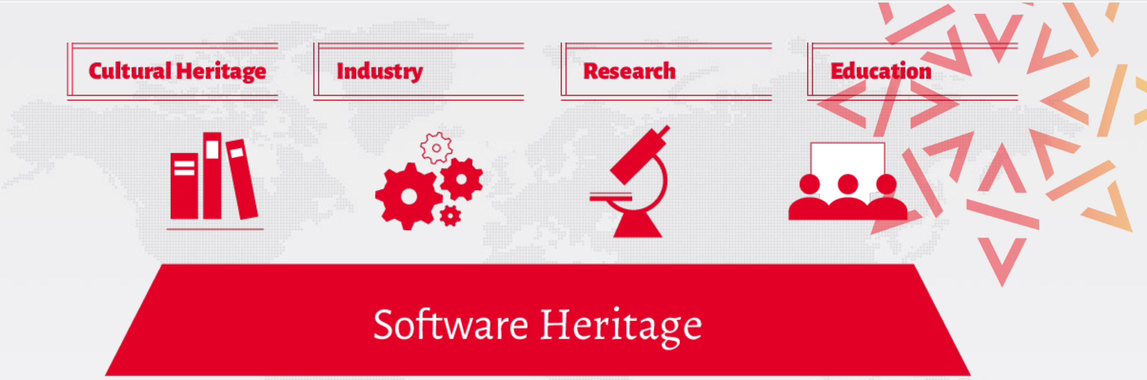
Industry

Research

Education

Software Heritage

Core principles



Cultural Heritage

Industry

Research

Education

Software Heritage

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

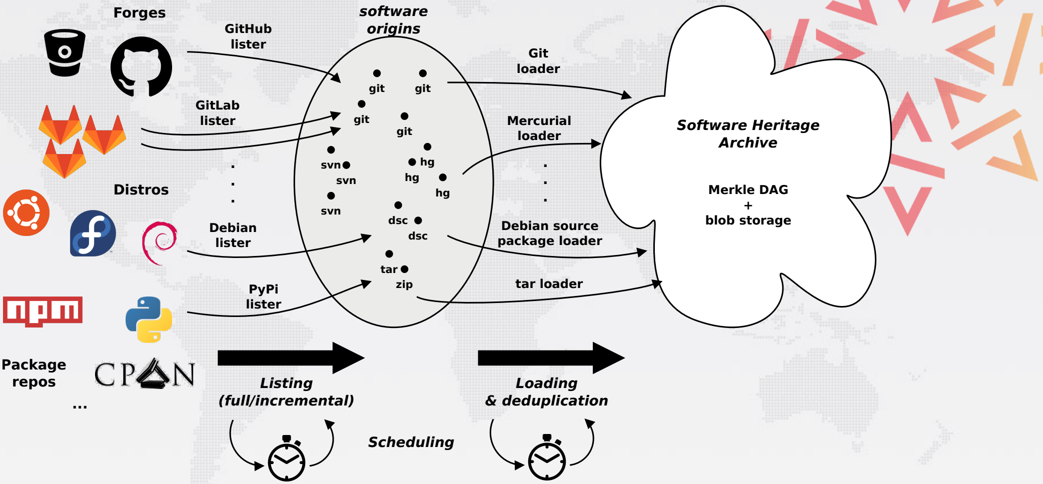
… in a VCS-/archive-agnostic canonical data model

We DON'T archive

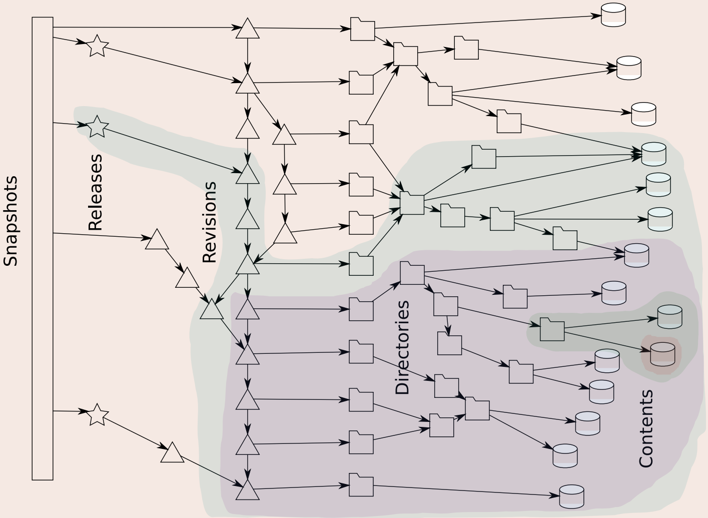
- homepages, wikis
- BTS/issues/code reviews/etc.
- mailing lists

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

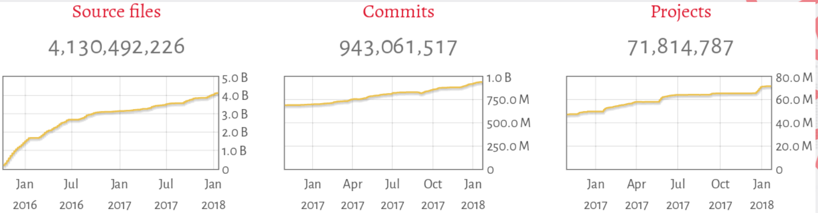
Data flow


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

The archive: a (giant) Merkle DAG



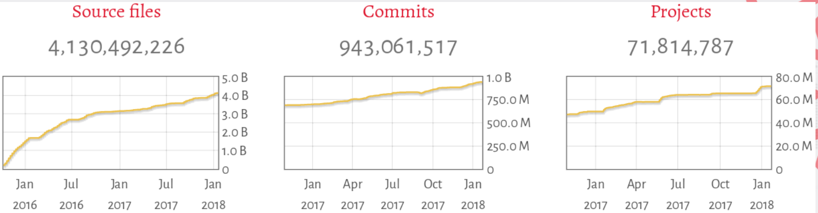
Archive coverage



Current sources

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

Archive coverage



Source files

4,130,492,226

Commits

943,061,517

Projects

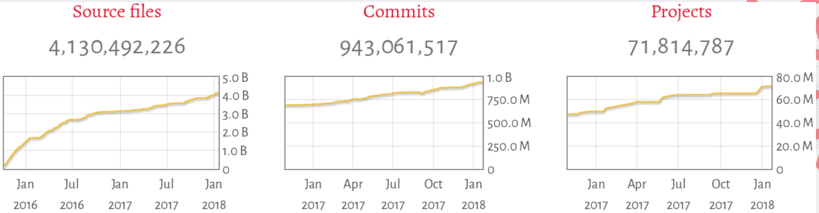
71,814,787

Current sources

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

150 TB blobs, 5 TB database (as a graph: 7 B nodes + 60 B edges)

Archive coverage



Current sources

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

150 TB blobs, 5 TB database (as a graph: 7 B nodes + 60 B edges)

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

Outline

1. The Free Software commons
2. Software Heritage
3. Accessing the archive
4. Getting involved

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/`

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"
}
```

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"
}
```

Caveats

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

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`

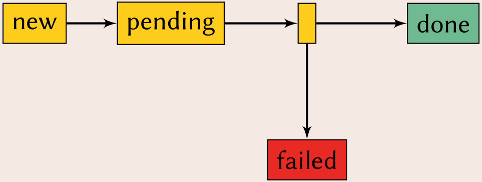
Request cooking

```
$ curl -X POST /api/1/vault/revision/a86747d2.../gitfast

{
  ’fetch_url’: ’/api/1/vault/revision/a86747d2.../gitfast/raw/’,
  ’progress_message’: None,
  ’status’: ’new’,
  ’id’: 4,
  ’obj_id’: ’a86747d201ab8f8657d145df4376676d5e47cf9f’,
  ’obj_type’: ’revision_gitfast’
}
```

Email notification

an optional email POST parameter can be used to request notification of bundle availability

Cooking progress



```
$ curl /api/1/vault/revision/a86747d2.../gitfast

{
  'fetch_url': '/api/1/vault/revision/a86747d2.../gitfast/raw/',
  'progress_message': None,
  'status': 'pending',
  'id': 4,
  'obj_id': 'a86747d201ab8f8657d145df4376676d5e47cf9f',
  'obj_type': 'revision_gitfast'
}
```

Download

```
$ curl /api/1/vault/revision/a86747d2.../gitfast

{
  ’fetch_url’: ’/api/1/vault/revision/a86747d2.../gitfast/raw/’,
  ’progress_message’: None,
  ’status’: ’done’,
  ’id’: 4,
  ’obj_id’: ’a86747d201ab8f8657d145df4376676d5e47cf9f’,
  ’obj_type’: ’revision_gitfast’
}
```

Download

```
$ curl /api/1/vault/revision/a86747d2.../gitfast

{
  ’fetch_url’: ’/api/1/vault/revision/a86747d2.../gitfast/raw/’,
  ’progress_message’: None,
  ’status’: ’done’,
  ’id’: 4,
  ’obj_id’: ’a86747d201ab8f8657d145df4376676d5e47cf9f’,
  ’obj_type’: ’revision_gitfast’
}
```

```
$ curl /api/1/vault/revision/a86747d2.../gitfast/raw/ \
  -O path/to/revision.gitfast.gz

$ git init
$ zcat path/to/revision.gitfast.gz | git fast-import
$ git checkout HEAD
```

Web user interface

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

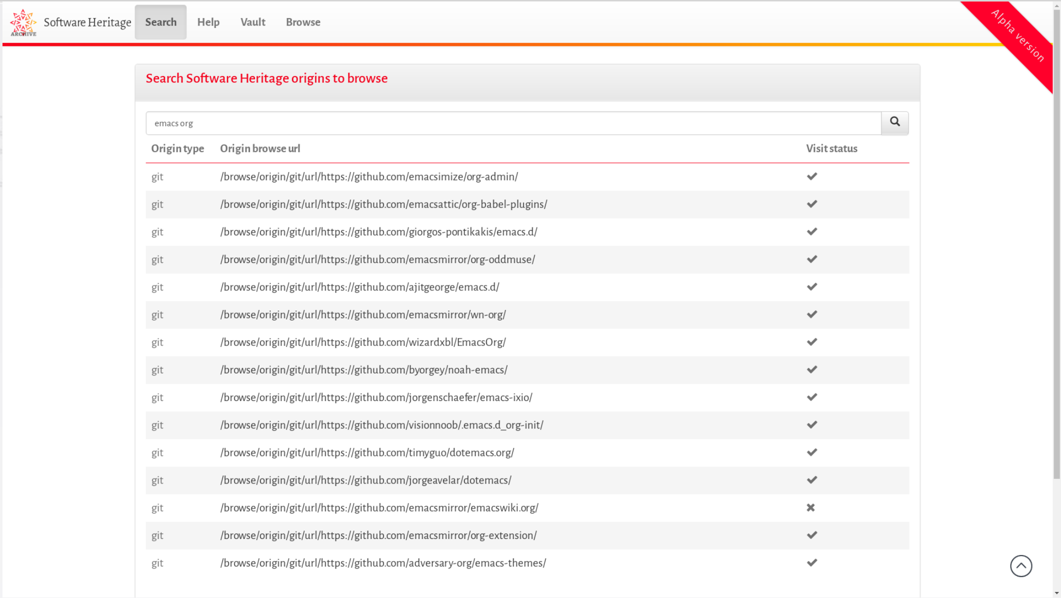
Technology preview… just for you!

- username: libreplanet
- password: 2018

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

Web UI — origin search

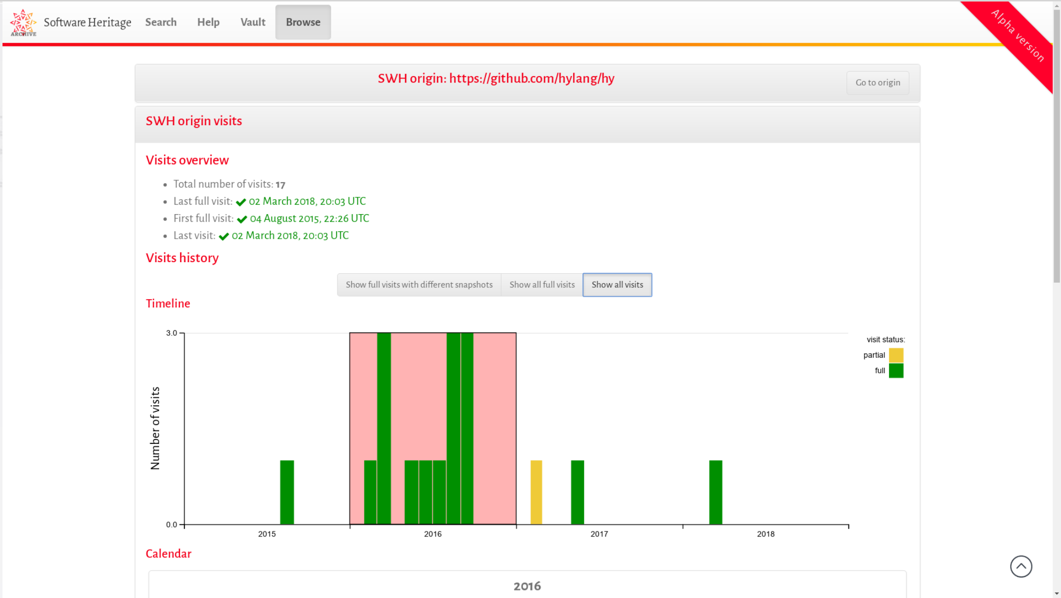


Search Software Heritage origins to browse

emacs org

Origin type	Origin browse url	Visit status
git	/browse/origin/git/url/https://github.com/emacsimize/org-admin/	✓
git	/browse/origin/git/url/https://github.com/emacsattic/org-babel-plugins/	✓
git	/browse/origin/git/url/https://github.com/giorgos-pontikakis/emacs.d/	✓
git	/browse/origin/git/url/https://github.com/emacsmirror/org-oddmuse/	✓
git	/browse/origin/git/url/https://github.com/ajitgeorge/emacs.d/	✓
git	/browse/origin/git/url/https://github.com/emacsmirror/wn-org/	✓
git	/browse/origin/git/url/https://github.com/wizardxbl/EmacsOrg/	✓
git	/browse/origin/git/url/https://github.com/byorgey/noah-emacs/	✓
git	/browse/origin/git/url/https://github.com/jorgenschaefer/emacs-ixio/	✓
git	/browse/origin/git/url/https://github.com/visionnoob/.emacs.d_org-init/	✓
git	/browse/origin/git/url/https://github.com/timyguo/dotemacs.org/	✓
git	/browse/origin/git/url/https://github.com/jorgeavelar/dotemacs/	✓
git	/browse/origin/git/url/https://github.com/emacsmirror/emacswiki.org/	✗
git	/browse/origin/git/url/https://github.com/emacsmirror/org-extension/	✓
git	/browse/origin/git/url/https://github.com/adversary-org/emacs-themes/	✓

Web UI — available visits



SWH origin: https://github.com/hylang/hy

SWH origin visits

Visits overview

- Total number of visits: **17**
- Last full visit: ✓ 02 March 2018, 20:03 UTC
- First full visit: ✓ 04 August 2015, 22:26 UTC
- Last visit: ✓ 02 March 2018, 20:03 UTC

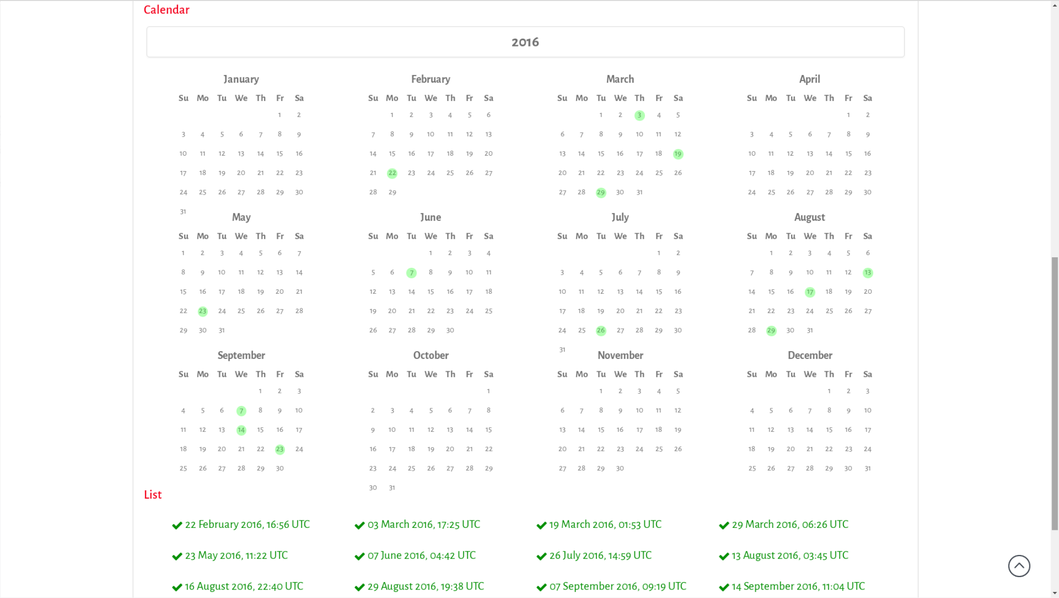
Visits history

Timeline

Calendar

2016

Web UI — calendar



Calendar

2016

January

Su	Mo	Tu	We	Th	Fr	Sa
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

February

Su	Mo	Tu	We	Th	Fr	Sa
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29					

March

Su	Mo	Tu	We	Th	Fr	Sa
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

April

Su	Mo	Tu	We	Th	Fr	Sa
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

May

Su	Mo	Tu	We	Th	Fr	Sa
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

June

Su	Mo	Tu	We	Th	Fr	Sa
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

July

Su	Mo	Tu	We	Th	Fr	Sa
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

August

Su	Mo	Tu	We	Th	Fr	Sa
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

September

Su	Mo	Tu	We	Th	Fr	Sa
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

October

Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

November

Su	Mo	Tu	We	Th	Fr	Sa
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			

December

Su	Mo	Tu	We	Th	Fr	Sa
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

List

✓ 22 February 2016, 16:56 UTC

✓ 03 March 2016, 17:25 UTC

✓ 19 March 2016, 01:53 UTC

✓ 29 March 2016, 06:26 UTC

✓ 23 May 2016, 11:22 UTC

✓ 07 June 2016, 04:42 UTC

✓ 26 July 2016, 14:59 UTC

✓ 13 August 2016, 03:45 UTC

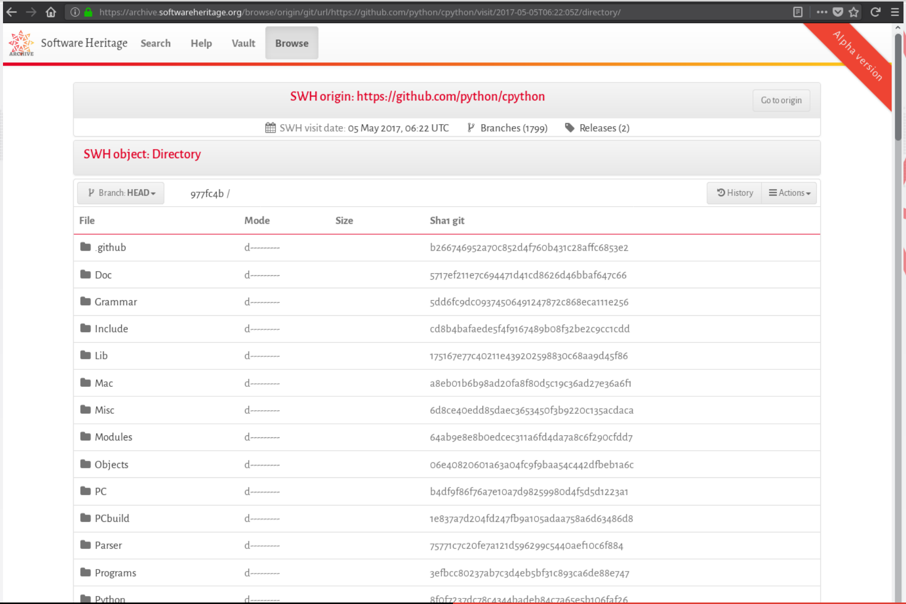
✓ 16 August 2016, 22:40 UTC

✓ 29 August 2016, 19:38 UTC

✓ 07 September 2016, 09:19 UTC

✓ 14 September 2016, 11:04 UTC

Web UI — directory browsing



https://archive.softwareheritage.org/browse/origin/git/url/https://github.com/python/cpython/visit/2017-05-05T06:22:05Z/directory/
Software Heritage
ARCHIVE
Search
Help
Vault
Browse
Alpha version
Go to origin
Branch: HEAD
History
Actions

SWH origin: https://github.com/python/cpython

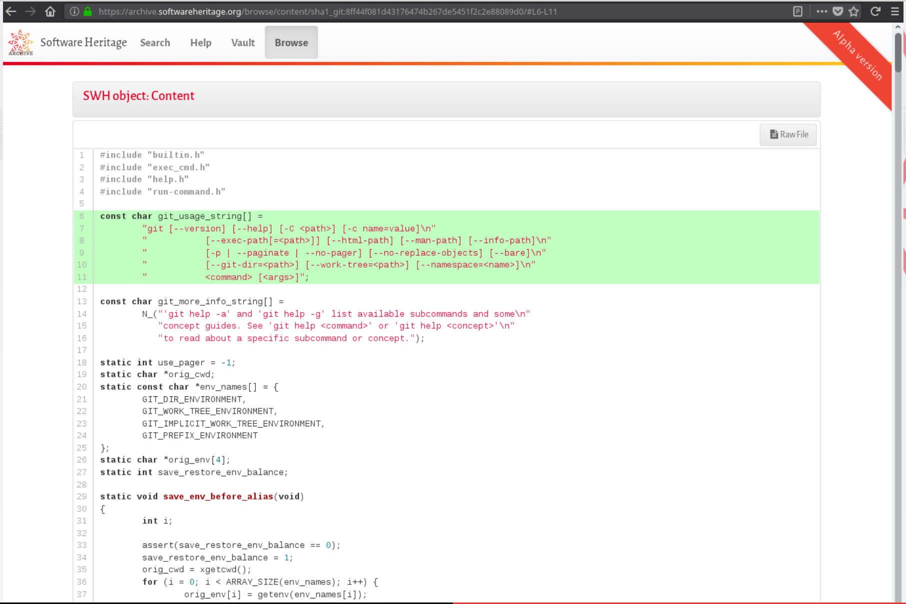
SWH visit date: 05 May 2017, 06:22 UTC   Branches (1799)   Releases (2)

SWH object: Directory

977fc4b /

File	Mode	Size	Sha1 git
.github	d---------		b266746952a70c852d4f760b431c28affc6853e2
Doc	d---------		5717ef211e7c694471d41cd8626d46bbaf647c66
Grammar	d---------		5dd6fc9dc0937450649124787 2c868eca111e256
Include	d---------		cd8b4bafaede5f4f9167489b08f32be2c9cc1cdd
Lib	d---------		175167e77c40211e439202598830c68aa9d45f86
Mac	d---------		a8eb01b6b98ad20fa8f80d5c19c36ad27e36a6f1
Misc	d---------		6d8ce40edd85daec3653450f3b9220c135acdaca
Modules	d---------		64ab9e8e8b0edcec311a6fd4da7a8c6f290cfdd7
Objects	d---------		06e40820601a63a04fc9f9baa54c442dfbeb1a6c
PC	d---------		b4df9f86f76a7e10a7d98259980d4f5d5d1223a1
PCbuild	d---------		1e837a7d204fd247fb9a105adaa758a6d63486d8
Parser	d---------		75771c7c20fe7a121d596299c5440aef10c6f884
Programs	d---------		3efbcc80237ab7c3d4eb5bf31c893ca6de88e747
Python	d---------		8f0f7237dc78c4344badeb84c7a65e5b106faf26

Web UI — syntax highlighting and selection

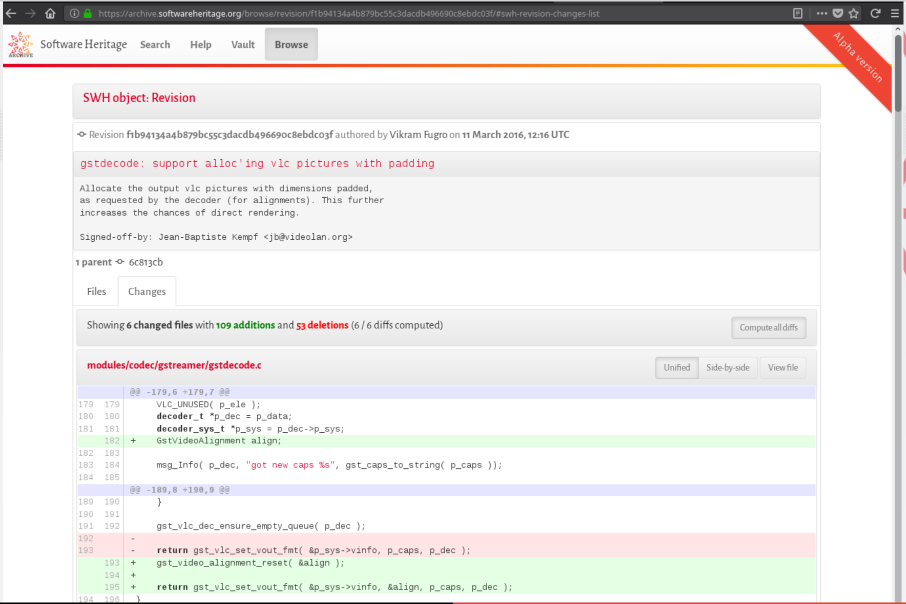


https://archive.softwareheritage.org/browse/content/sha1_git:8ff44f081d43176474b267de5451f2c2e88089d0/#L6-L11
Software Heritage
ARCHIVE
Search
Help
Vault
Browse
Alpha version
Raw File

SWH object: Content

```
#include "builtin.h"
#include "exec_cmd.h"
#include "help.h"
#include "run-command.h"

const char git_usage_string[] =
        "git [--version] [--help] [-C <path>] [-c name=value]\n"
        "           [--exec-path[=<path>]] [--html-path] [--man-path] [--info-path]\n"
        "           [-p | --paginate | --no-pager] [--no-replace-objects] [--bare]\n"
        "           [--git-dir=<path>] [--work-tree=<path>] [--namespace=<name>]\n"
        "           <command> [<args>]";

const char git_more_info_string[] =
        N_("'git help -a' and 'git help -g' list available subcommands and some\n"
           "concept guides. See 'git help <command>' or 'git help <concept>'\n"
           "to read about a specific subcommand or concept.");

static int use_pager = -1;
static char *orig_cwd;
static const char *env_names[] = {
        GIT_DIR_ENVIRONMENT,
        GIT_WORK_TREE_ENVIRONMENT,
        GIT_IMPLICIT_WORK_TREE_ENVIRONMENT,
        GIT_PREFIX_ENVIRONMENT
};
static char *orig_env[4];
static int save_restore_env_balance;

static void save_env_before_alias(void)
{
        int i;

        assert(save_restore_env_balance == 0);
        save_restore_env_balance = 1;
        orig_cwd = xgetcwd();
        for (i = 0; i < ARRAY_SIZE(env_names); i++) {
                orig_env[i] = getenv(env_names[i]);
```

Web UI — revisions as diffs


https://archive.softwareheritage.org/browse/revision/f1b94134a4b879bc55c3dacdb496690c8ebdc03f/#swh-revision-changes-list
Software Heritage
ARCHIVE
Search
Help
Vault
Browse
Alpha version
Files
Changes
Compute all diffs
Unified
Side-by-side
View file

SWH object: Revision

Revision **f1b94134a4b879bc55c3dacdb496690c8ebdc03f** authored by Vikram Fugro on **11 March 2016, 12:16 UTC**

```
gstdecode: support alloc'ing vlc pictures with padding

Allocate the output vlc pictures with dimensions padded,
as requested by the decoder (for alignments). This further
increases the chances of direct rendering.

Signed-off-by: Jean-Baptiste Kempf <jb@videolan.org>
```

1 parent 6c813cb

Showing **6 changed files** with **109 additions** and **53 deletions** (6 / 6 diffs computed)

modules/codec/gstreamer/gstdecode.c

```
@@ -179,6 +179,7 @@
     VLC_UNUSED( p_ele );
     decoder_t *p_dec = p_data;
     decoder_sys_t *p_sys = p_dec->p_sys;
+    GstVideoAlignment align;

     msg_Info( p_dec, "got new caps %s", gst_caps_to_string( p_caps ));

@@ -189,8 +190,9 @@
     }

     gst_vlc_dec_ensure_empty_queue( p_dec );
-
-    return gst_vlc_set_vout_fmt( &p_sys->vinfo, p_caps, p_dec );
+    gst_video_alignment_reset( &align );
+
+    return gst_vlc_set_vout_fmt( &p_sys->vinfo, &align, p_caps, p_dec );
 }
```

Outline

1. The Free Software commons
2. Software Heritage
3. Accessing the archive
4. Getting involved

Roadmap

Features…

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

Roadmap

Features…

- (done) lookup by content hash
- browsing: "wayback machine" for archived code
  - (done) via Web API
  - (early access) via Web UI
- (early access) deposit of source code bundles directly to the archive
- (early access) download: `wget` / `git clone` from the archive
- (todo) provenance lookup 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 at hand's reach!

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

Software Heritage is

- a reference archive of all Free Software ever written
- an international, open, nonprofit, mutualized infrastructure
- now accessible to developers, users, vendors
- at the service of our community, at the service of society

Come in, we're open!

www.softwareheritage.org - general information
wiki.softwareheritage.org - internships, leads
forge.softwareheritage.org - our own code

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