

# Software Heritage

## The Great Library of Source Code

Nicolas Dandrimont

Software Engineer - Software Heritage  
olasd@softwareheritage.org

26 october 2019  
MiniDebConf Vaumarcus



# Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

# Software is everywhere





Software embodies a growing part of...

... our scientific, *technical* and Cultural Heritage!

# Source Code: *executable* and *human readable* knowledge



*“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 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");
}
```

## Apollo 11 (~60.000 lines), 1969



"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton





# ~ 50 years, a lightning fast growth

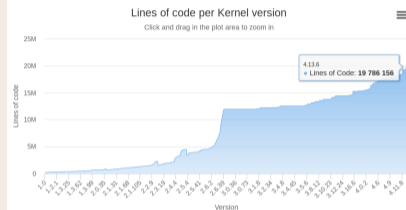
## Apollo 11 (~60.000 lines), 1969



"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton

## Linux Kernel (in your pockets!)



~ 50 years, a lightning fast growth

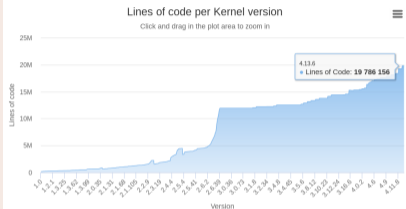
Apollo 11 (~60.000 lines), 1969



"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton

Linux Kernel (in your pockets!)



Harold Abelson, Structure and Interpretation of Computer Programs

(1985)

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

# ~ 50 years, a lightning fast growth

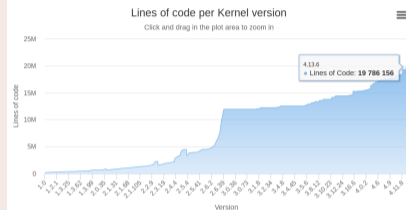
Apollo 11 (~60.000 lines), 1969



"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton

Linux Kernel (in your pockets!)



Harold Abelson, Structure and Interpretation of Computer Programs (1985)

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

Len Shustek, Computer History Museum (2006)

*"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, dangling, wear, corruption, encryption, format, attack, aging, media, tear, obsolete, and dependencies. The words are arranged in a circular pattern with varying colors and sizes.

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 centered on a faint world map background. The words are in various colors and orientations, including: damage, disaster, malicious, deletion, reference, storage, dangling, wear, corruption, encryption, format, attack, aging, media, tear, obsolete, and dependencies.

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?

# We are at a turning point

Preserve the past

Only a few years left to recover the history of software technology

# We are at a turning point

## Preserve the past

Only a few years left to recover the history of software technology

## Improve the future

We need a **universal** platform for all the future software 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.*



**Cultural Heritage**



**Industry**



**Research**



**Education**



**Software Heritage**

**Cultural Heritage**



**Industry**



**Research**



**Education**



## Software Heritage

### Open approach

- 100% Free and Open Source 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

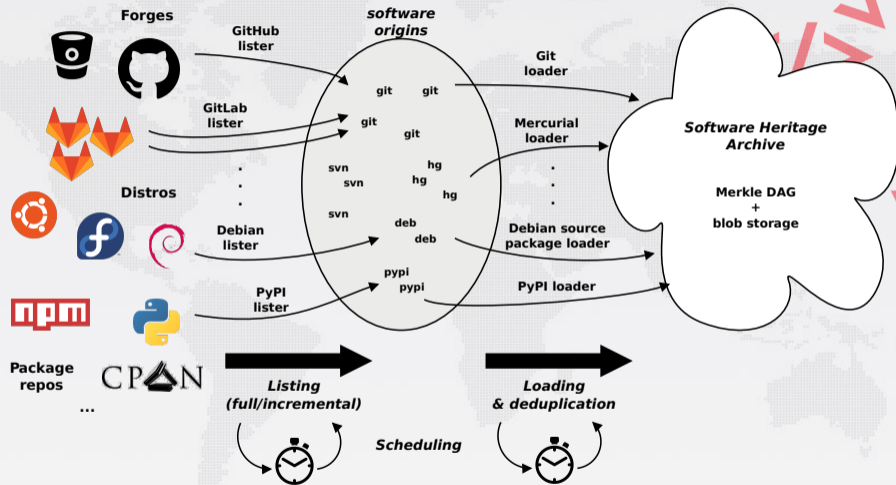
... in a VCS-/archive-agnostic **canonical data model**

## We DON'T archive

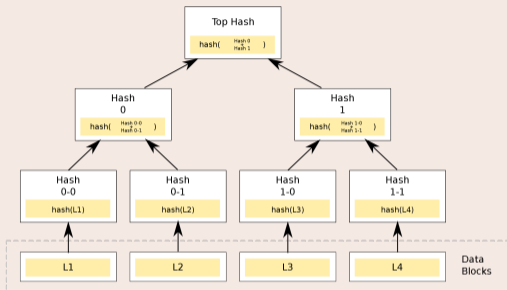
- homepages, wikis
- BTS/issues/code reviews/etc.
- mailing lists

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

# Data flow



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

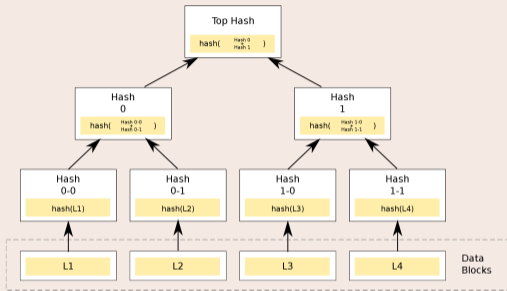


Combination of

- tree
- hash function

# Merkle trees

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



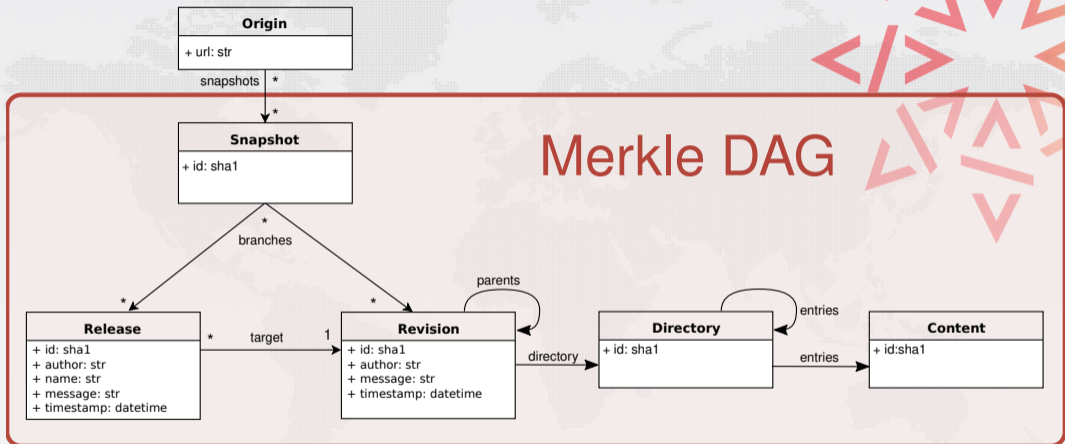
Combination of

- 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



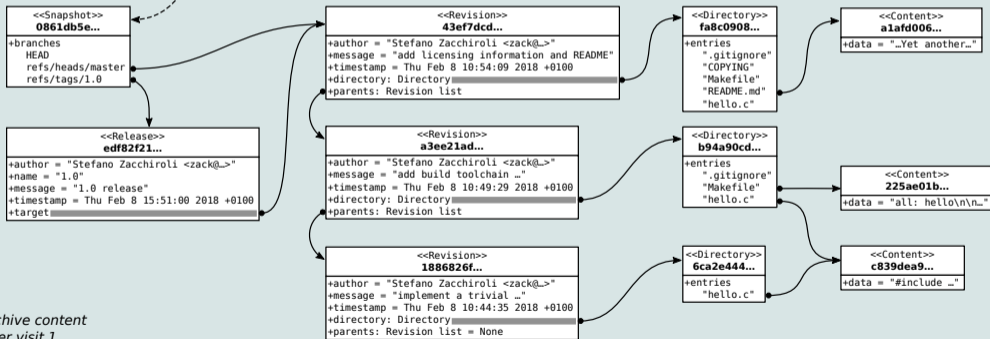
# The archive: a (giant) Merkle DAG

origin  
https://forge.softwareheritage.org/source/helloworld.git

visit  
1

snapshot  
0861db5e...

timestamp  
Fri Feb 9 12:38:45 2018 +0100

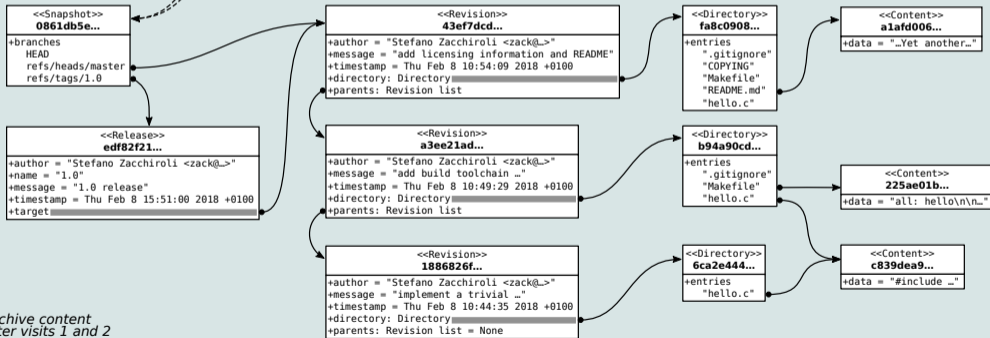


Archive content  
after visit 1



# 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 Zacchiroli <zack@...>"
+name = "1.0"
+message = "1.0 release"
+timestamp = Thu Feb 8 15:51:00 2018 +0100
+target
```

```
<<Revision>>
c7640e8d...
+author = "Stefano Zacchiroli <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 Zacchiroli <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 Zacchiroli <zack@...>"
+message = "add build toolchain ..."
+timestamp = Thu Feb 8 10:49:29 2018 +0100
+directory: Directory
+parents: Revision list
```

```
<<Revision>>
1886826f...
+author = "Stefano Zacchiroli <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..."
```

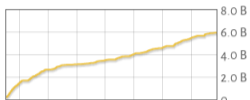
```
<<Content>>
225ae01b...
+data = "all: hello\n\n_"
```

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

Archive content  
after visits 1 and 2

## Source files

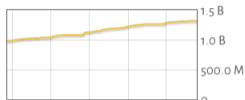
6,006,503,960



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

## Commits

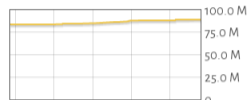
1,326,776,432



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

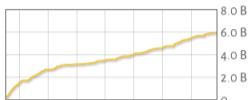
archives-ouvertes.fr

Inria  
inventeurs du monde numérique



## Source files

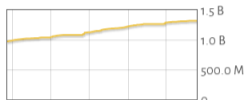
6,006,503,960



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

## Commits

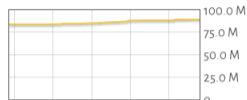
1,326,776,432



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

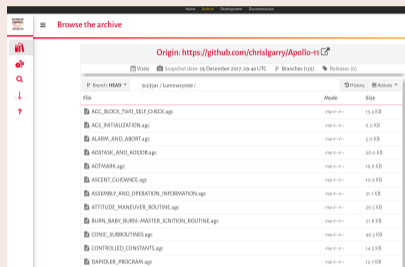
- all **REST API features**, but good looking :-)
  - browsing: snapshots → revisions → directories → contents ...
  - access to metadata and crawling information
- **origin search**, as full text indexing of origin URLs
- bulk **download**, via integration with the Vault

# Visiting the archive: the Apollo 11 source code

## Margaret Hamilton today



## The Apollo 11 source code in SWH



Origin: <https://github.com/chrisgarry/Apollo-11>

File	Mode	Size
AGC_BLOCK_TWO_SELF_CHECK.agc	rw-r--r--	15.4 KB
AGC_INITIALIZATION.agc	rw-r--r--	5.0 KB
ALARM_AND_ABORT.agc	rw-r--r--	5.0 KB
ADSTAR_AND_ADSDOB.agc	rw-r--r--	28.0 KB
AOTMARK.agc	rw-r--r--	18.8 KB
ASCENT_GUIDANCE.agc	rw-r--r--	10.9 KB
ASSEMBLY_AND_OPERATION_INFORMATION.agc	rw-r--r--	31.1 KB
ATTITUDE_MANEUVER_ROUTINE.agc	rw-r--r--	26.5 KB
BURN_BABY_BURN-MASTER_ICNITION_ROUTINE.agc	rw-r--r--	21.8 KB
CDNIC_SUBROUTINES.agc	rw-r--r--	45.5 KB
CONTROLLED_CONSTANTS.agc	rw-r--r--	14.5 KB
DAPIOLER_PROGRAM.agc	rw-r--r--	12.7 KB

## Some pointers

- Entry point
- Burn, baby, burn!

# Visiting the archive: the Quake 3 source code

## John Carmack



## The Quake 3 source code in SWH

A screenshot of the Software Heritage (SWH) archive interface. The page title is "Browse the archive". The origin is listed as "https://github.com/id-Software/Quake-III-Arena". The snapshot date is "23 October 2017, 12:24 UTC". The current branch is "HEAD" with commit "c5f0212". A table lists the files and directories in the archive:

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

Below the table, there is a red link for "README.txt" and a red heading "Quake III Arena GPL source release".

## Some pointers

- Entry point
- What the f...



## Search Software Heritage origins to browse

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/wm-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/orgenschaefer/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/orgavelar/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/	✓







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

Go to origin

## 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

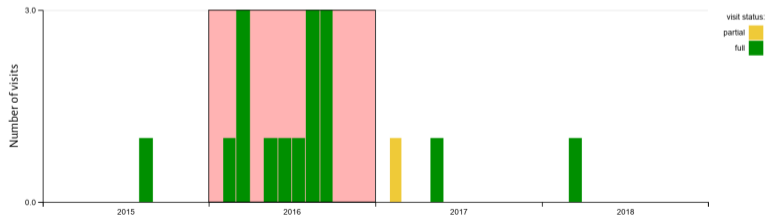
### Visits history

Show full visits with different snapshots

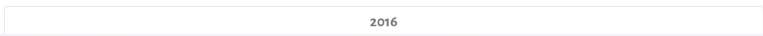
Show all full visits

Show all visits

### Timeline



### 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						

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				

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

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

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

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					

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					

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					

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						

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						

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					

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

✓ 23 May 2016, 11:22 UTC

✓ 16 August 2016, 22:40 UTC

✓ 03 March 2016, 17:25 UTC

✓ 07 June 2016, 04:42 UTC

✓ 29 August 2016, 19:38 UTC

✓ 19 March 2016, 01:53 UTC

✓ 26 July 2016, 14:59 UTC

✓ 07 September 2016, 09:19 UTC

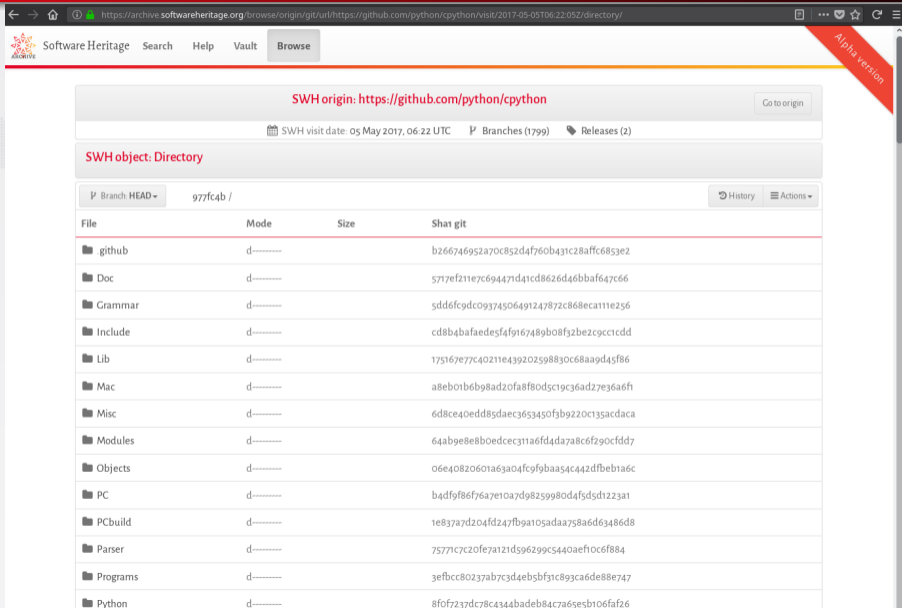
✓ 29 March 2016, 06:26 UTC

✓ 13 August 2016, 03:45 UTC

✓ 14 September 2016, 11:04 UTC



# Web UI — directory browsing



The screenshot shows the Software Heritage web interface for a directory. At the top, there's a navigation bar with 'Software Heritage', 'Search', 'Help', 'Vault', and 'Browse' buttons. Below that, a header indicates the SWH origin as 'https://github.com/python/cpython' with a 'Go to origin' button. A secondary header shows the SWH visit date as '05 May 2017, 06:22 UTC' and statistics for 'Branches (1799)' and 'Releases (2)'. The main content area shows the SWH object as 'Directory' and the current branch as 'HEAD - 977fc4b /'. Below this is a table listing files and directories with columns for 'File', 'Mode', 'Size', and 'Sha1 git'. A red banner in the top right corner says 'Alpha version'.

SWH origin: <https://github.com/python/cpython> Go to origin

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

SWH object: Directory

Branch: HEAD - 977fc4b / History Actions

File	Mode	Size	Sha1 git
github	d-----		b266746952a70c852d4f760b431c28affc6853e2
Doc	d-----		5717ef211e7c694471d41cd8626d46bbaf647c66
Grammar	d-----		5dd6fc9dc09374506491247872c868eca111e256
Include	d-----		cd8b4bafaedesf4f9167489b08f32be2c9cc1cdd
Lib	d-----		175167e77c40211e439202598830c68aa9d45f86
Mac	d-----		a8eb01b6b98ad20fa8f80d5c19c36ad27e36a6f1
Misc	d-----		6d8ce40edd85daec3653450f3b9220c135acdaca
Modules	d-----		64ab9e8e8boedcec311a6fd4da7a8c6f290cfd7
Objects	d-----		06e40820601a63a04fc9f9baa54c442dfb6b1a6c
PC	d-----		b4df986f76a7e10a7d98259980d4f5d5d1223a1
PCbuild	d-----		1e837a7d204fd247fb9a105adaa758a6d63486d8
Parser	d-----		75771c7c20fe7a121d596299c5440aef10c6f884
Programs	d-----		3efbcc80237ab7c3d4eb5bf31c893ca6de88e747
Python	d-----		8f0f7237dc78c4344badeb84c7a65esb106faf26

# Web UI – syntax highlighting and selection

Software Heritage Search Help Vault Browse

SWH object: Content Raw File

```
1 #include "builtin.h"
2 #include "exec_cmd.h"
3 #include "help.h"
4 #include "run-command.h"
5
6 const char git_usage_string[] =
7     "git [--version] [--help] [-C <path>] [-c name=value]\n"
8     "    [--exec-path=<path>] [--html-path] [--man-path] [--info-path]\n"
9     "    [-p | --paginate | --no-pager] [--no-replace-objects] [--bare]\n"
10    "    [--git-dir=<path>] [--work-tree=<path>] [--namespace=<name>]\n"
11    "    <command> [<args>]";
12
13 const char git_more_info_string[] =
14     N_("\"git help -a\" and 'git help -g' list available subcommands and some\n"
15     "concept guides. See 'git help <command>' or 'git help <concept>'\n"
16     "to read about a specific subcommand or concept.");
17
18 static int use_pager = -1;
19 static char *orig_cwd;
20 static const char *env_names[] = {
21     GIT_DIR_ENVIRONMENT,
22     GIT_WORK_TREE_ENVIRONMENT,
23     GIT_IMPLICIT_WORK_TREE_ENVIRONMENT,
24     GIT_PREFIX_ENVIRONMENT
25 };
26 static char *orig_env[4];
27 static int save_restore_env_balance;
28
29 static void save_env_before_alias(void)
30 {
31     int i;
32
33     assert(save_restore_env_balance == 0);
34     save_restore_env_balance = 1;
35     orig_cwd = xgetcwd();
36     for (i = 0; i < ARRAY_SIZE(env_names); i++) {
37         orig_env[i] = getenv(env_names[i]);
38         orig_env[i] = xstrdup_or_null(orig_env[i]);
39     }
40 }
```

# Web UI — revisions as diffs

Software Heritage Search Help Vault Browse

Alpha version

**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

Files Changes

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

**modules/codecs/gstreamer/gstdecode.c** Unified Side-by-side View file

```
@@ -179,6 +179,7 @@
179 179     VLC_UNUSED( p_ele );
180 180     decoder_t *p_dec = p_data;
181 181     decoder_sys_t *p_sys = p_dec->p_sys;
182 +   GstVideoAlignment align;
182 183
183 184     msg_Info( p_dec, "got new caps %s", gst_caps_to_string( p_caps ));
184 185
@@ -189,8 +190,9 @@
189 190     }
190 191
191 192     gst_vlc_dec_ensure_empty_queue( p_dec );
192 -
193 +   return gst_vlc_set_vout_fmt( &p_sys->vinfo, p_caps, p_dec );
193 +   gst_video_alignment_reset( &align );
194 +
195 +   return gst_vlc_set_vout_fmt( &p_sys->vinfo, &align, p_caps, p_dec );
194 196 }
```

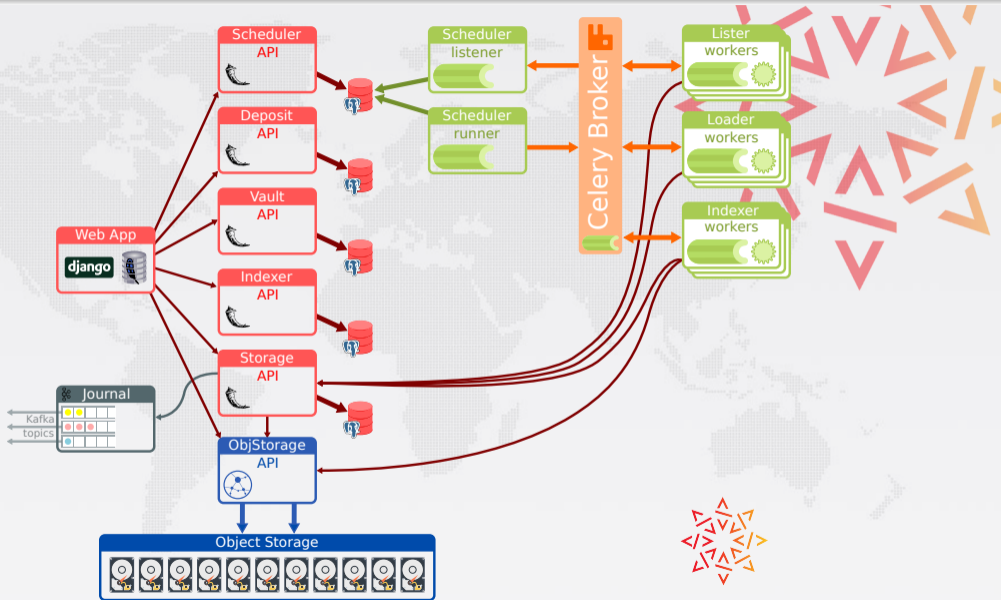
## 3rd party

- Debian, Proxmox, ZFS on Linux, Puppet
- 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)

# Deployment architecture



## in house

- 3x hypervisors with ~40 VMs
- 1x high performance database server; read-only replica on a container
- 2x dedicated storage servers, one of them using ZFS.
- 3x high density storage array (2 x 60 x 6TB; 1 x 60 x 10TB)
- 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/>

## Increase archive coverage - Archit Agrawal

- Listens for more upstream repositories (Ruby, NodeJS, CRAN, GNU Project)
- Refactoring of loaders for simpler implementation of new instances (e.g. CPAN?)

## End-to-end testing of the web frontend - Kalpit Kothari

- using cypress
- automates a lot of clicky QA work

## Graph compression - Thibault Allançon

- Handle hundreds of billions of edges in (sensibly sized) memory
- Fast algorithms for large scale analysis... on a single machine
- Toolkit for source code analysis by the scientific community

## Mirror infrastructure

- Based on kafka, and a dedicated Python component
- Pushes all object types to topics as soon as they're added
- Allows building a full replica of the archive, or just subscribing to some object types

## Mirroring agreements

- Two partner organizations hosting a full mirror (Work in Progress)
- FOSSID
- ENEA

## Paris Call: Software Source Code as Heritage for Sustainable Development

- Signed in February 2019, by a group of experts on Software Preservation, with the support of UNESCO
- Global awareness of the importance of preserving software in general, and software source code in particular, for sustainable development.

## Software Heritage Acquisition Protocol

- Published in October 2019, in collaboration with the University of Pisa
- Step by step Collection and Curation of meaningful historical Software Source Code

# Yes, now you can!

Wayback machine for source code

retrieve the source code as it was

Open science

deposit scientific software (via HAL)

Reference catalog

use intrinsic identifiers for software

Universal knowledge base

store the knowledge about source code

And much, much more is in store!

With your help?

## The Software Heritage Foundation

- independent
- long term mission
- multistakeholder

## The community

- academia: Open Access, research
- industry: better software
- cultural heritage: **all** the software history

## The mirror network

- resilience
- biodiversity

*“Let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.”*

*Thomas Jefferson*



# Software Heritage

www.softwareheritage.org

@swheritage

Everybody is concerned, everybody can help build

## The Library of Alexandria of code



- recover the past
- structure the future

## A CERN for Software



- build better software
  - for industry
  - for society as a whole

# 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