Software Heritage
The Great Library of Source Code

Nicolas Dandrimont
Software Engineer - Software Heritage
olasd@softwareheritage.org

26 October 2019
MiniDebConf Vaumarcus
Software is everywhere

Software embodies a growing part of our scientific, technical and Cultural Heritage!
Software is everywhere

Software embodies a growing part of…

… our scientific, *technical* and Cultural Heritage!
“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.”

Hello World

Source Code: executable and human readable knowledge

GPL Licence

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

**Hello World**

<table>
<thead>
<tr>
<th>Program (excerpt of binary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4004e6: 55</td>
</tr>
<tr>
<td>4004e7: 48 89 e5</td>
</tr>
<tr>
<td>4004ea: bf 84 05 40 00</td>
</tr>
<tr>
<td>4004ef: b8 00 00 00 00</td>
</tr>
<tr>
<td>4004f4: e8 c7 fe ff ff</td>
</tr>
<tr>
<td>4004f9: 90</td>
</tr>
<tr>
<td>4004fa: 5d</td>
</tr>
<tr>
<td>4004fb: c3</td>
</tr>
</tbody>
</table>

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

Hello World

Program (source code)

/* Hello World program */

#include<stdio.h>

void main()
{
    printf("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
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
"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."
Margaret Hamilton
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.”
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)
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
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

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 trees

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

Merkle DAG

- **Origin**
  - url: str

- **Snapshot**
  - id: sha1

- **Release**
  - id: sha1
  - author: str
  - name: str
  - message: str
  - timestamp: datetime

- **Revision**
  - id: sha1
  - author: str
  - message: str
  - timestamp: datetime

- **Directory**
  - id: sha1
  - entries

- **Content**
  - id: sha1

- **parents**
  - branches

- **target**
  - 1

- **entries**

Nicolas Dandrimont
The archive: a (giant) Merkle DAG

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

visit
1
snapshot
0861db5e...
Fri Feb 9 12:38:45 2018 +0100
2
0861db5e...
Fri Feb 9 13:29:00 2018 +0100

timestamp
Fri Feb 9 12:38:45 2018 +0100
Fri Feb 9 13:29:00 2018 +0100

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

<<Content>>
ea1afd006...
+data = "...Yet another..."

<<Content>>
fa8c0908...
+data = "...Yet another..."

<<Content>>
0861db5e...

<<Directory>>
b94a90cd...
+entries
  .gitignore
  COPYING
  Makefile
  README.md
  hello.c
+data = "all: hello\n"

<<Directory>>
fa8c0908...
+entries
  .gitignore
  COPYING
  Makefile
  README.md
  hello.c
+data = "all: hello\n"

<<Directory>>
6ca2e444...
+entries
  hello.c
+data = "#include ...

<<Directory>>
fa8c0908...
+entries
  .gitignore
  COPYING
  Makefile
  README.md
  hello.c
+data = "all: hello\n"

<<Directory>>
b94a90cd...
+entries
  .gitignore
  COPYING
  Makefile
  README.md
  hello.c
+data = "all: hello\n"

<<Directory>>
1886826f...
+entries
  .gitignore
  COPYING
  Makefile
  README.md
  hello.c
+data = "#include ...

<<Snapshot>>
0861db5e...

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

archive content after visits 1 and 2

Nicolas Dandrimont

Software Heritage 2019-10-26 Vaumarcus 14 / 35
The archive: a (giant) Merkle DAG

Archive content after visits 1 and 2

Archive content after visits 1, 2 and 3
Archive coverage — archive.softwareheritage.org

Source files
6,006,503,960

Commits
1,326,776,432

Projects
89,301,694

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

The richest public source code archive, ... and growing daily!
Web user interface

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

Some pointers

- Entry point
- Burn, baby, burn!
Visiting the archive: the Quake 3 source code

John Carmack

The Quake 3 source code in SWH

Browse the archive

Origin: https://github.com/id-Software/Quake-III-Arena

<table>
<thead>
<tr>
<th>File</th>
<th>Mode</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>code</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>common</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>sys</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>lex</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>include</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>script</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>standard</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>src</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>COPYING.txt</td>
<td></td>
<td>12.8K</td>
</tr>
<tr>
<td>README.txt</td>
<td></td>
<td>0.0K</td>
</tr>
</tbody>
</table>

Quake III Arena GPL source release

Some pointers

- Entry point
- What the f…
**Web UI — origin search**

<table>
<thead>
<tr>
<th>Origin type</th>
<th>Origin browse url</th>
<th>Visit status</th>
</tr>
</thead>
<tbody>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/emacsimize/org-admin/">https://github.com/emacsimize/org-admin/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/emacsattic/org-babel-plugins/">https://github.com/emacsattic/org-babel-plugins/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/giorgos-pontikakis/emacs.d/">https://github.com/giorgos-pontikakis/emacs.d/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/emacs-pontikakis/emacs.d/">https://github.com/emacs-pontikakis/emacs.d/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/ajitgeorge/emacs.d/">https://github.com/ajitgeorge/emacs.d/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/emacs/score/wn-org/">https://github.com/emacs/score/wn-org/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/wizardxbl/EmacsOrg/">https://github.com/wizardxbl/EmacsOrg/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/byorgey/noah-emacs/">https://github.com/byorgey/noah-emacs/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/jorgenschaefler/emacs-ixio/">https://github.com/jorgenschaefler/emacs-ixio/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/visionnoob/emacs.d_org-init/">https://github.com/visionnoob/emacs.d_org-init/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/timyguo/dotemacs.org/">https://github.com/timyguo/dotemacs.org/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/jorgeavelar/dotemacs/">https://github.com/jorgeavelar/dotemacs/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/emacs-minor-emacs-wiki.org">https://github.com/emacs-minor-emacs-wiki.org</a></td>
<td>✗</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/emacs-minor/org-extension/">https://github.com/emacs-minor/org-extension/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/adversary-org/emacs-themes/">https://github.com/adversary-org/emacs-themes/</a></td>
<td>✔</td>
</tr>
</tbody>
</table>
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

Number of visits

Calendar
Web UI — directory browsing

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

SWH object: Directory

Branch: HEAD /

File          Mode    Size    Sha git
---           ------   ------   ----------
github        d--------  b26674952a70c852d4f760b43128aff6653e2
Doc           d--------  5775ef2119769447d410d8626d46bba6647c66
Grammar       d--------  5df66f9d09374564f1272c868cfc11c756
Include       d--------  cd88b8baed5efcf97e47f79b08f2b2e29c1cddd
Lib           d--------  1751677240211431920598830d8a9d4d8c86
Mac           d--------  afebe002a7d5ad2f68ff08d9c1935d27c33a0f1
Misc          d--------  6d8a02e6d8d1663435f9b3220131a2dca
Modules       d--------  6ab9f8e8e8b0ed1c31a6d5a7c0d6290cfd7
Objects       d--------  064e0e2b604fa63a13f9f885f442dfebe8a6c
PC            d--------  b045f86f67a4c1c2d2d825598041f5d5f6623a1
PCBuild       d--------  10837ad20ffd2af7b9a105a0a5386d63468d8
Parser        d--------  75777c7c07f6742795e5d92605340d11005f884
Programs     d--------  3efb8c80237a17d5e5eb5f3c83ca06e8e8e247
Python        d--------  8f0f572f6d7bc4344badb847add5e5eb166fa26

Nicolas Dandrimont
include "builtin.h"
#include "exec_cmd.h"
#include "help.h"
#include "run-command.h"

const char git_usage_string[] =
  "git [-version] [-help] [-C <path>] [-o <name=value>\n"
  "  [--exec-path=<path>] [--exec-path[<path>]]\n"
  "  [--html-path] [--macute-path] [--info-path] \n"
  "  [--page-file] [--no-pager] [--no-replace-objects] [--bare]\n"
  "  [--git-dir=<path>] [--work-tree=<path>] [--namespace=<name>] \n"
  "  <command> [arguments];";

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_MUPVIGIT_WORK_TREE_ENVIRONMENT,
  GIT_SPARSE_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 = xgotcwd();
  for (i = 0; i < ARRAY_SIZE(env_names); i++) {
    orig_env[i] = getenv(env_names[i]);
    if (!orig_env[i]) orig_env[i] = strdup("error or pull/copy.env(i)");
    orig_env[i] = xputenv(env_names[i], orig_env[i]);
    if (!orig_env[i]) orig_env[i] = strdup("error or pull/copy.env(i)");
  }
Web UI — revisions as diffs

---

**SWH object: Revision**

- Revision fb9413a44b879bc55d9adb9e6690c8ebd103f authored by Vikram Fugro on 11 March 2016, 12:16 UTC

**gstdecode: support allocating 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 <jkb@videolan.org>

Files

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

```
modules/code/gstreamer/gstdecode.c

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;
182 182  *gstvideoaclignment align;
183 183  msg_Info( p_dec, "got new caps %s", gst_caps_to_string( p_caps ));
184 185
190 190  gst_vlc_dec_unprepare_queue( p_dec );
192 192  return gst_vlc_set_vout_fmt( &p_sys->info, p_caps, p_dec );
193 193  + gst_video_alignment_reset( &align );
194 194  + return gst_vlc_set_vout_fmt( &p_sys->info, &align, p_caps, p_dec );
```

---
Software stack

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)
Hardware stack

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
Software development

**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/
Google Summer of Code

Increase archive coverage - Archit Agrawal
- Listers 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
Replication

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

Reference catalog
- use intrinsic identifiers for software

Open science
- deposit scientific software (via HAL)

Universal knowledge base
- store the knowledge about source code

And much, much more is in store!

With your help?
The next steps

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
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

Nicolas Dandrimont
Software Heritage 2019-10-26 Vaumarcus 35 / 35
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