Software Heritage
The Great Library of Source Code

David Douard & Antoine R. Dumont
Software Engineer - Software Heritage
info@softwareheritage.org

27 juin 2019
PSES - Choisy-le-Roy
Software is everywhere
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.”

David Douard & Antoine R. Dumont

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

Hello World

GPL Licence

“Hello World” Program (excerpt of binary)

```assembly
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.”

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."

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

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

Software Heritage
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"
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

Origin
+ url: str

Snapshots
+ id: str

Release
+ id: str
+ author: str
+ name: str
+ message: str
+ timestamp: datetime

Revision
+ id: str
+ author: str
+ message: str
+ timestamp: datetime

Directory
+ id: str

Content
+ sha1: str

Merkle DAG

Origin
+ url: str
branches
Snapshot
+ id: str

Release
+ id: str
branches 1
Revision
parents
Directory
entries
Content
entries
+ sha1: str

David Douard & Antoine R. Dumont
Software Heritage 2019-06-27 PSES 12 / 32
The archive: a (giant) Merkle DAG

Archive content after visit 1

Visit

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

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

Snapshots

1

Timestamp

Fri Feb 9 12:38:45 2018 +0100

<<Revision>>

43ef7dcd...

<<Directory>>

b94a90cd...

<<Release>>

edf82f21...

<<Content>>

fa8c0908...

<<Revision>>

a3ee21ad...

<<Directory>>

fa8c0908...

<<Revision>>

1886826f...

<<Directory>>

6ca2e444...

<<Content>>

c839dea9...

<<Revision>>

0861db5e...

<<Directory>>

b94a90cd...

<<Revision>>

43ef7dcd...

<<Directory>>

b94a90cd...

<<Revision>>

a3ee21ad...

<<Directory>>

fa8c0908...

<<Revision>>

1886826f...

<<Directory>>

6ca2e444...

<<Revision>>

0861db5e...

<<Directory>>

b94a90cd...

<<Revision>>

a3ee21ad...

<<Directory>>

fa8c0908...

<<Revision>>

1886826f...

<<Directory>>

6ca2e444...

<<Revision>>

0861db5e...

<<Directory>>

b94a90cd...

<<Revision>>

a3ee21ad...

<<Directory>>

fa8c0908...

<<Revision>>

1886826f...

<<Directory>>

6ca2e444...

<<Revision>>

0861db5e...

<<Directory>>

b94a90cd...

<<Revision>>

a3ee21ad...

<<Directory>>

fa8c0908...

<<Revision>>

1886826f...

<<Directory>>

6ca2e444...
The archive: a (giant) Merkle DAG

Archive content after visits 1 and 2
The archive: a (giant) Merkle DAG

Archive content after visits 1 and 2

Archive content after visits 1, 2 and 3

Origin visits

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

Visit | Snapshot | Timestamp
--- | --- | ---
1 | 0861db5e... | Fri Feb 9 12:38:45 2018 +0100
2 | 0861db5e... | Fri Feb 9 13:29:00 2018 +0100
3 | 510aa88b... | Fri Feb 9 15:52:50 2018 +0100

Revision a3ee21ad...

+author = "Stefano Zacchiroli <zack@...>
+message = "move source code to src/
+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

Revision edf82f21...

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

Revision 45f0c078...

+author = "Stefano Zacchiroli <zack@...>
+message = "COPYING"
+timestamp = Thu Feb 8 15:26:08 2018 +0100
+directory: Directory
+parents: Revision list

Directory fa8c0908...

+entries "COPYING"
+entries "Makefile"
+entries "README.md"
+entries "hello.c"

Directory b94a90cd...

+entries ".gitignore"
+entries "COPYING"
+entries "Makefile"
+entries "README.md"
+entries "hello.c"

Directory 0861db5e...

+branches HEAD
+refs/heads/master
+refs/heads/doc
+refs/tags/1.0

Directory 510aa88b...

+branches HEAD
+refs/heads/master
+refs/tags/1.0

Directory a802e2f21...

+branches HEAD
+refs/heads/master
+refs/tags/1.0

Directory c839de902...

+entries "hello.c"
+entries ".gitignore"
+entries "COPYING"
+entries "Makefile"
+entries "README.md"
+entries "src"

Directory 510aa88b...

+entries "COPYING"
+entries "Makefile"
+entries "README.md"
+entries "src"

Directory 0861db5e...

+entries "COPYING"
+entries "Makefile"
+entries "README.md"
+entries "src"

Directory 510aa88b...

+parents: Revision list

Directory 0861db5e...

+parents: Revision list

Release edf82f21...

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

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

Snapshot 510aa88b...

+branches HEAD
+refs/heads/master
+refs/tags/1.0

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 = "move source code to src/
+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

Revision edf82f21...

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

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 = "move source code to src/
+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

Revision edf82f21...

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

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 = "move source code to src/
+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

Revision edf82f21...

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

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 = "move source code to src/
+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

Revision edf82f21...

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

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 = "move source code to src/
+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
Archive coverage — archive.softwareheritage.org

The richest public source code archive, and growing daily!

<table>
<thead>
<tr>
<th>Source files</th>
<th>Commits</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,006,503,960</td>
<td>1,326,776,432</td>
<td>89,301,694</td>
</tr>
</tbody>
</table>

GitHub, debian, GitLab, Google code, python, GITORIOUS, GNU, HAL, Inria
Archive coverage — archive.softwareheritage.org

- 410 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

Some pointers

- Entry point
- What the f…
### Search Software Heritage origins to browse

<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/emacsinside/org-admin/">https://github.com/emacsinside/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/igorgios-pontikakis/emacs.d/">https://github.com/igorgios-pontikakis/emacs.d/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/emacs-mirror/org-oddmuse/">https://github.com/emacs-mirror/org-oddmuse/</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-mirror/vim-org/">https://github.com/emacs-mirror/vim-org/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/wizard-box/EmacsOrg/">https://github.com/wizard-box/EmacsOrg/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/borgey/noah-emacs/">https://github.com/borgey/noah-emacs/</a></td>
<td>✔</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/jorgenschefer/emacs-ixio/">https://github.com/jorgenschefer/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/trimyguo/dotemacs.org/">https://github.com/trimyguo/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-mirror/emacswiki.org/">https://github.com/emacs-mirror/emacswiki.org/</a></td>
<td>❌</td>
</tr>
<tr>
<td>git</td>
<td>/browse/origin/git/url/<a href="https://github.com/emacs-mirror/org-extension/">https://github.com/emacs-mirror/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>
Calendar

2016

January

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

February

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

March

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

April

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

May

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

June

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

July

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

August

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

September

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

October

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

November

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

December

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

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

David Douard & Antoine R. Dumont

Software Heritage 2019-06-27 PSES
Web UI — directory browsing

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

SWH object: Directory

<table>
<thead>
<tr>
<th>File</th>
<th>Mode</th>
<th>Size</th>
<th>Sha1 git</th>
</tr>
</thead>
<tbody>
<tr>
<td>github</td>
<td>d----</td>
<td></td>
<td>b26674952a70c852da4f760ba43c28a4fc68353ed2</td>
</tr>
<tr>
<td>Doc</td>
<td>d----</td>
<td></td>
<td>577ef21167654471d410d8626d46baef647c66</td>
</tr>
<tr>
<td>Grammar</td>
<td>d----</td>
<td></td>
<td>5dd6f00c0957a5a06a7f272c868ec1c112236</td>
</tr>
<tr>
<td>Include</td>
<td>d----</td>
<td></td>
<td>cd6b2b8a3ed52f8f574bdc0bf8ebe29c1cddd</td>
</tr>
<tr>
<td>Lib</td>
<td>d----</td>
<td></td>
<td>17516777c42111e43202259830dc68a99d45f86</td>
</tr>
<tr>
<td>Mac</td>
<td>d----</td>
<td></td>
<td>a8eb01b6b9ad2a0a8800d4c19c3b27756d669f1</td>
</tr>
<tr>
<td>Misc</td>
<td>d----</td>
<td></td>
<td>6d38c40cdd5a85e65c40c83b22c0c13ac0ca6a</td>
</tr>
<tr>
<td>Modules</td>
<td>d----</td>
<td></td>
<td>6a9b2e8e8b0eddec311aef4d3a7a8c659c3c0f7f7</td>
</tr>
<tr>
<td>Objects</td>
<td>d----</td>
<td></td>
<td>0e4b82b601a6a0a0f5y4bba3443dfb3ba6f1</td>
</tr>
<tr>
<td>PC</td>
<td>d----</td>
<td></td>
<td>b3df5f867670c07d825550b7c4f5d5d122321</td>
</tr>
<tr>
<td>PCBuild</td>
<td>d----</td>
<td></td>
<td>1833f2b7a2d4f7f9b9150da5a278a6ed6346d4</td>
</tr>
<tr>
<td>Parser</td>
<td>d----</td>
<td></td>
<td>75777cc20f7221d962395449aef0c68f884</td>
</tr>
<tr>
<td>Programs</td>
<td>d----</td>
<td></td>
<td>3ebf33c8237a37dc4ebeb5f38c93c8e6e867a7</td>
</tr>
<tr>
<td>Python</td>
<td>d----</td>
<td></td>
<td>4557b2c20f4c7a2a49d43b2f38c93c8e6e867a7</td>
</tr>
</tbody>
</table>
Web UI — syntax highlighting and selection

```c
#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] [-n | --no-pager] [-r | --repository]\n"
  " [-R | --replace-objects] [-B | --bare]\n"
  " [-g | --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 usa_pager = -1;
static char *orig_cwd;
static const char *env_names[] = {
    GIT_DIR_ENVIRONMENT,
    GIT_WORK_TREE_ENVIRONMENT,
    GIT_MIRROR_WORK_TREE_ENVIRONMENT,
    GIT_PREFERRED_ENVIRONMENT
};
static char *orig_env[4];
static int save_restore_env_balance;
static void save_restore_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] = gatom(env_names[i]);
    }
}
```
Web UI — revisions as diffs

SWH object: Revision

Revision: fb9413a4ab1b871b9c55c1daedbb9e66d01c8ebd0c3f authored by Vikram Rugro 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 <jhb@videolan.org>

t parent: 6c813cb

Showing 6 changed files with 109 additions and 39 deletions (6/6 diffs computed)

Diff of modules/codec/gststreamer/gstdecode.c

```c
179 179  VLC戡MSED( p_e;e );
180 180  decoder_t *p_dec = p_data;
181 181  decoder_sys_t *p_sys = p_dec->p_sys;
182+ 182+  gstdimeAlignment aligned;
183 183  msg_Info( p_dec, 'got new caps %a', gstd_caps_to_string( p_caps ));
184 184
185+ 185+  VLC戡 videocpDec_ensure_empty_queue( p_dec );
186 186  }
190 190  gstd_video_alignment_reset( &align );
194+ 194+  return gstd_video_set_vout_fmt( &p_sys->xinfo, &p_caps, p_dec );
195+ 195+  return gstd_video_set_vout_fmt( &p_sys->xinfo, &p_caps, p_dec );
```

So/f_tware stack

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)
## Hardware stack

### 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
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/
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
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
Reproducibility is the key

non-reproducible single occurrences are of no significance to science

Karl Popper, The Logic of Scientific Discovery, 1934

the software source code is special

- It embodies the logic of the data transformation.
- It must be included in strategies for scientific knowledge preservation.
- knowing the exact version of the software used is essential for reproducibility.
Deposit software in HAL

http://hal.inria.fr/hal-01738741

How to do it:
- today: deposit .zip or .tar.gz file (guide)
- tomorrow: provide SWH id and metadata
  include metadata file for automatic metadata extraction

September 2018: open to all on https://hal.archives-ouvertes.fr/

Click on the links in the paper and view the source code!

“Our Parmap.parmap and Parmap.parfold functions may be used to seamlessly . . .”

All features of So/f_tware Heritage for free

intrinsic IDs (integrity, not dependent on resolvers!)

browse, download (now)
metadata, licenses, provenance (plagiarism detection), classification (wip), . . .

Coverage and uniformity
one archive for all domains (industry included)
reference any so/f_tware, not just the deposited ones
git-compatible identifiers greatly simplify workflows

Sustainability . . . doors are open!

one infrastructure
independent non profit foundation
worldwide
mirrors
Deposit software in HAL

http://hal.inria.fr/hal-01738741

Generic mechanism:
- SWORD based
- review process
- versioning

How to do it:
today: deposit .zip or .tar.gz file
tomorrow: provide SWH id and metadata
include metadata file for automatic metadata extraction

September 2018: open to all on https://hal.archives-ouvertes.fr/
Click on the links in the paper and view the source code!

“Our Parmap.parmap and Parmap.parfold functions may be used to seamlessly . . . ”

All features of So/f_tware Heritage for free
intrinsic IDs (integrity, not dependent on resolvers!)
browse, download (now)
metadata, licenses, provenance (plagiarism detection), classification (wip), . . .

Coverage and uniformity
one archive for all domains (industry included)
reference any so/f_tware, not just the deposited ones

Sustainability . . . doors are open!

One infrastructure independent non profit foundation worldwide
mirrors
Deposit software in HAL

http://hal.inria.fr/hal-01738741

Generic mechanism:
- SWORD based
- review process
- versioning

How to do it:
- today: deposit .zip or .tar.gz file (guide)
- tomorrow:
  - provide SWH id and metadata
  - include metadata file for automatic metadata extraction
  - ...

September 2018: open to all on https://hal.archives-ouvertes.fr/

Click on the links in the paper and view the source code!

“Our Parmap.parmap and Parmap.parfold functions may be used to seamlessly . . .”

All features of Software Heritage for free
intrinsic IDs (integrity, not dependent on resolvers!)
browse, download (now)
metadata, licenses, provenance (plagiarism detection), classification (wip), . . .

Coverage and uniformity
one archive for all domains (industry included)
reference any software, not just the deposited ones
git-compatible identifiers greatly simplify workflows
Sustainability . . . doors are open!

one infrastructure
independent
non-profit foundation
worldwide mirrors
Deposit software in HAL

http://hal.inria.fr/hal-01738741

Generic mechanism:
- SWORD based
- review process
- versioning

How to do it:
- today: deposit .zip or .tar.gz file (guide)
- tomorrow:
  - provide SWH id and metadata
  - include metadata file for automatic metadata extraction
  - ...

September 2018: open to all on https://hal.archives-ouvertes.fr/
Deposit software in HAL

http://hal.inria.fr/hal-01738741

Generic mechanism:
- SWORD based
- review process
- versioning

How to do it:
- today: deposit .zip or .tar.gz file (guide)
- tomorrow:
  - provide SWH id and metadata
  - include metadata file for automatic metadata extraction
  - ...

September 2018: open to all on https://hal.archives-ouvertes.fr/
Deposit software in HAL

http://hal.inria.fr/hal-01738741

Generic mechanism:
- SWORD based
- review process
- versioning

How to do it:
- **today**: deposit .zip or .tar.gz file (**guide**)
- **tomorrow**:
  - provide **SWH id** and metadata
  - include *metadata file* for automatic metadata extraction
  - ...

September 2018: **open to all** on https://hal.archives-ouvertes.fr/