Software Heritage

Why and How We Preserve all of Mankind's Source Code

Roberto Di Cosmo

roberto@dicosmo.org

April 20th, 2018

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Software all around us
2. The Software Heritage initiative
3. Building for the long term
4. Conclusion

Software is everywhere

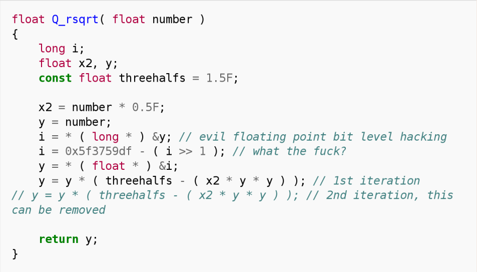
Software embodies our collective Knowledge and Cultural Heritage

Software Source Code is *special*

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

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

Quake 2 source code (excerpt)

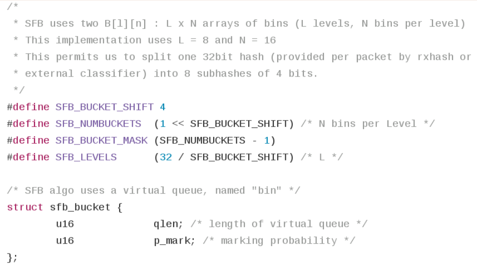


```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    x2 = number * 0.5F;
    y = number;
    i = * ( long * ) &y; // evil floating point bit level hacking
    i = 0x5f3759df - ( i >> 1 ); // what the fuck?
    y = * ( float * ) &i;
    y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
// y = y * ( threehalfs - ( x2 * y * y ) ); // 2nd iteration, this
can be removed

    return y;
}
```

Net. queue in Linux (excerpt)



```
/*
 * SFB uses two B[l][n] : L x N arrays of bins (L levels, N bins per level)
 * This implementation uses L = 8 and N = 16
 * This permits us to split one 32bit hash (provided per packet by rxhash or
 * external classifier) into 8 subhashes of 4 bits.
 */
#define SFB_BUCKET_SHIFT 4
#define SFB_NUMBUCKETS  (1 << SFB_BUCKET_SHIFT) /* N bins per Level */
#define SFB_BUCKET_MASK (SFB_NUMBUCKETS - 1)
#define SFB_LEVELS      (32 / SFB_BUCKET_SHIFT) /* L */

/* SFB algo uses a virtual queue, named "bin" */
struct sfb_bucket {
        u16             qlen; /* length of virtual queue */
        u16             p_mark; /* marking probability */
};
```

Len Shustek, Computer History Museum

"Source code provides a view into the mind of the designer."

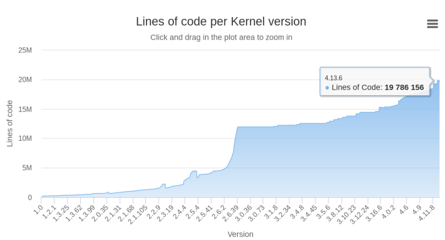
~ 50 years, a lightning fast growth

Apollo 11 Guidance Computer (~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



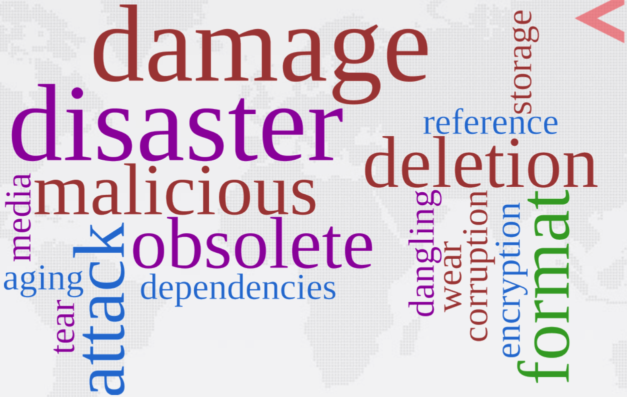
… now in your pockets!

are we taking care of all this?

Software is spread all around

Sourceforge
Debian
CPAN
Gitorious
Maven
Bitbucket
Inria
GitHub
BerliOs
CTAN
Gitlab
CRAN
GoogleCode
Adullact

Software is fragile



Software lacks its own research infrastructure

Photo: ALMA(ESO/NAOJ/NRAO), R. Hills

Outline

1. Software all around us

2. The Software Heritage initiative

3. Building for the long term

4. Conclusion

The Software Heritage Project

Software Heritage

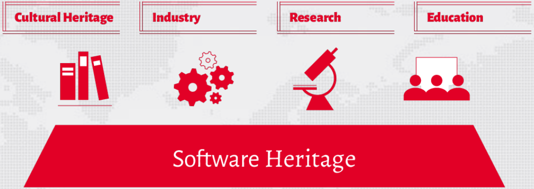
Our mission

Collect, preserve and share the *source code* of *all the software* that is available

Past, present and future

Preserving the past, *enhancing* the present, *preparing* the future

A principled infrastructure http://bit.ly/swhpaper



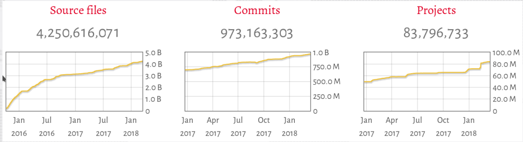
Cultural Heritage

Industry

Research

Education

Software Heritage



Source files

4,250,616,071

Commits

973,163,303

Projects

83,796,733

Technology

- transparency and FOSS
- replicas all the way down

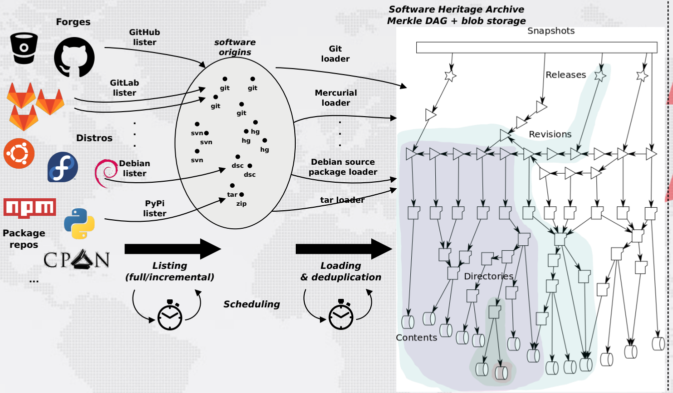
Content

- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- multi-stakeholder

Architecture (simplified)



- full development history permanently archived
- origins: GitHub (automated), Debian (automated), Gitorious, Google Code, GNU
- ~150Tb raw contents, ~10Tb graph (7+Bn nodes, 60+Bn edges)

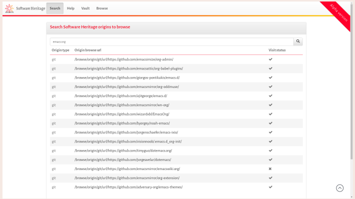
Using the archive

Preview *just for you*: a "wayback machine" for archived code!

- go to http://archive.softwareheritage.org/browse
- use the credentials: devoxx / 2018

No time for a demo, let's highlight some features. . .

Origin search

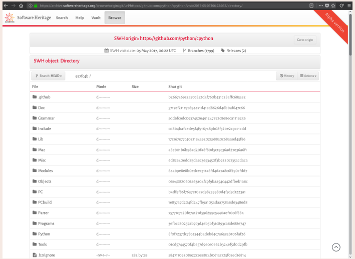


Search Software Heritage origins to browse

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/	✓

Directory browsing



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

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

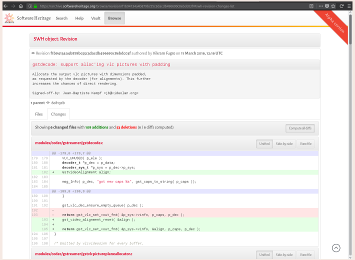
SWH visit date: 05 May 2017, 06:22 UTC   Branches (1799)   Releases (2)

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
Tools	d---------		01cd5744570f4be57d9ece0e62b324ef5d0d29fb
.bzrignore	-rw-r--r--	582 bytes	58471109208922c9ee8c4b06135725f03ed16814

Revisions as diffs



https://archive.softwareheritage.org/browse/revision/f1b94134a4b879bc55c3dacdb496690c8ebdc03f/#swh-revision-changes-list
Software Heritage
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 );
  }

  /* Emitted by vlcvideosink for every buffer,
```

modules/codec/gstreamer/gstvlcpictureplaneallocator.c

Outline

1. Software all around us

2. The Software Heritage initiative

3. Building for the long term

4. Conclusion

Growing Support

Landmark Inria Unesco agreement, April 3rd, 2017

Inria
INVENTEURS DU MONDE NUMÉRIQUE

UNESCO

Sharing the vision

IEEE Software · GitLab · ENGINEERING · eclipse · OW2 · fsfe · nlnet FOUNDATION · SIGSOFT SPECIAL INTEREST GROUP ON SOFTWARE ENGINEERING · Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE · INAPI · THE LINUX FOUNDATION · ADULLACT · INFORMATICS EUROPE · Computer History Museum · software freedom conservancy · Software Freedom Law Center · AdaCore The GNAT Pro Company · gandi.net · TDF · FREE SOFTWARE FOUNDATION · creative commons · SIF Société Informatique de France · open source initiative · openinventionnetwork

Contributing to the mission

Inria
INVENTEURS DU MONDE NUMÉRIQUE

>= 100Ke/year	Microsoft · intel · SOCIETE GENERALE
>= 50Ke/year	
>= 25Ke/year	HUAWEI
>= 10Ke/year	Data Archiving and Networked Services DANS · NOKIA Bell Labs · ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA DIPARTIMENTO DI INFORMATICA - SCIENZA E INGEGNERIA · GitHub

You can help!

Coding, advice

see forge.softwareheritage.org

Current development priorities ... *all* contributions equally welcome!

★★★	documentation
★★	listers/loaders for unsupported forges, VCS
★★	Web UI improvements

Yes, we'll be hiring, see www.softwareheritage.org/jobs

Funding

- make *your company* a sponsor : sponsorship.softwareheritage.org
- give *your own contribution* : www.softwareheritage.org/donate

Spread the word!

- follow and relay project news
- share the vision, tell others how to support the mission

Outline

1. Software all around us
2. The Software Heritage initiative
3. Building for the long term
4. Conclusion

Come in, we're open!

Software Heritage

www.softwareheritage.org @swheritage

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