Software Heritage

Building the Universal Software Archive for Open Science

Roberto Di Cosmo

roberto@dicosmo.org

September 19th, 2018

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

Source code is *special*

Harold Abelson, Structure and Interpretation of Computer Programs

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

Quake III 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."

Outline

Software is spread all around

Software is fragile

Software lacks its own research infrastructure

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

Outline

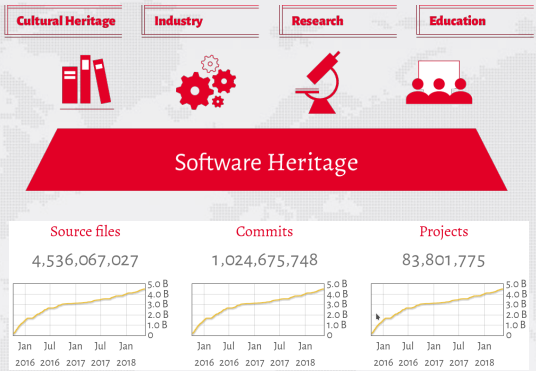
Software Heritage

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



Technology

- transparency and FOSS
- replicas all the way down

Content

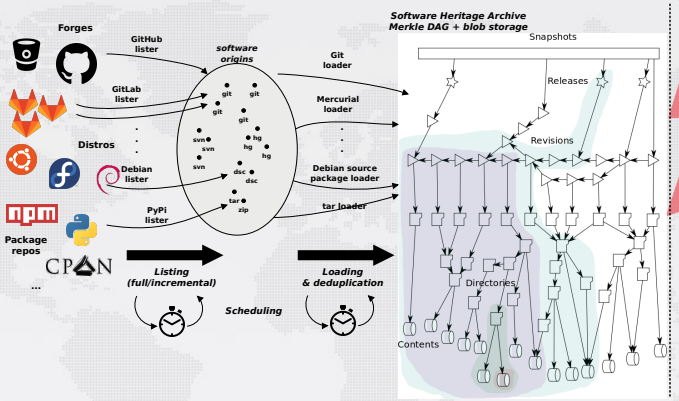
- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- mirror network

Outline

Automation, and storage



- full development history permanently archived
- origins: GitHub (auto), Debian (auto), Gitlab.com, Gitorious, Google Code, GNU
- ~ 200Tb raw contents, ~ 10Tb graph (10Bn nodes, 100Bn edges)

Outline

Research software: a long way to go!

ICSE (Zannier, Melrik, Maurer, 2006)

- complete absence of replication studies

ACM TOSEM 2001 to 2006 C. Ghezzi `http://bit.ly/tosemreprod`

- 60% of all papers have tools: only 20% *installable*

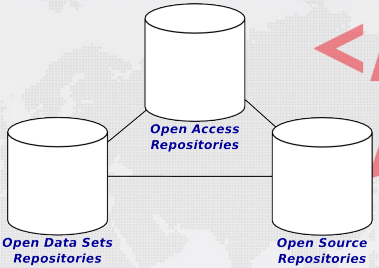
Collberg's 2015 study `http://reproducibility.cs.arizona.edu/`

- 601 mainstream papers: 508 with tools, only 40% *installable*

Main reasons

source code (*or the right version of it*) cannot be found

Supporting more accessible and reproducible science



A global library referencing all software used in all research fields

- completes the infrastructure for Open Access in science
- provides intrinsic persistent identifiers for scientific reproducibility
- enables large scale, verifiable software studies

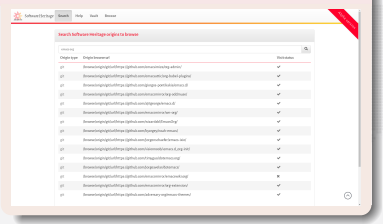
Reference archive for all software

A "wayback machine" for software source code … with intrinsic identifiers!

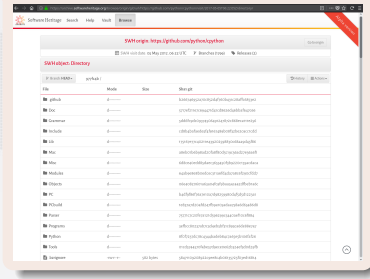
- http://archive.softwareheritage.org/browse
- http://bit.ly/swhpids for persistent identifiers

Demo time: let's highlight some features…

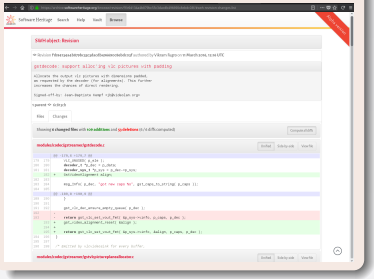
Origin search



Directory browsing

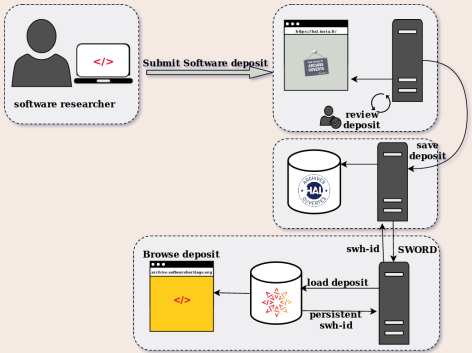


Revisions as diffs



Deposit Scientific Software

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/

The way to go to archive and reference scientific software

All features of Software Heritage *for free*

- intrinsic IDs (integrity, not dependent on resolvers!)
  - specification: `http://bit.ly/swhpids`
  - iPres2018 paper: `http://bit.ly/swhpidpaper`
- 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

Outline

Big Code = Big data + AI

Large scale *repeatable* software studies…

- vulnerability detection
- dependency analysis
- pattern elicitation
- automatic classification …

… need a uniform representation

Software Heritage has one data model for all forges/VCS…

… yes, we do data normalization of software evolution!

Breaking news: *soon* an Amazon public data set!

Outline

Growing Support

Landmark Inria Unesco agreement, April 3rd, 2017

Inria — INVENTEURS DU MONDE NUMÉRIQUE

UNESCO

Sharing the vision

Software Heritage, GitLab, Engineering, eclipse, OW2, fsfe, nlnet Foundation, SIGSOFT, Liberté · Égalité · Fraternité République Française, INRIA, The Linux Foundation, ADULLACT, Informatics Europe, Computer History Museum, software freedom conservancy, Software Freedom Law Center, AdaCore, gandi.net, TDF, FREE SOFTWARE FOUNDATION, creative commons, SIF, open source initiative, openinventionnetwork

Contributing to the mission

Inria — INVENTEURS DU MONDE NUMÉRIQUE

>= 100Ke/year	Microsoft, intel, SOCIETE GENERALE
>= 50Ke/year	HUAWEI, Google
>= 25Ke/year	DANS, NOKIA Bell Labs, FOSSID
>= 10Ke/year	GitHub, UQÀM

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

You can help!

Many scientific and technological challenges

object storage, machine learning, classification, efficient graph queries, mirror protocols, ...

Contribute

- `forge.softwareheritage.org`

Funding

- become a partner/sponsor/mirror : `sponsorship.softwareheritage.org`
- give *your own contribution* : `www.softwareheritage.org/donate`

Spread the word!

- *use* the archive and help others do
- tell everybody about Software Heritage

Outline

Come in, we're open!

Software Heritage

www.softwareheritage.org @swheritage

Library of Alexandria of code

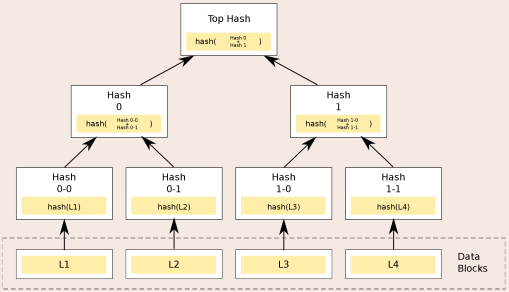
- recover the past
- structure the future

A CERN for Software

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

Much more than an archive!

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

Outline

URL decay disrupts the *web of reference*

Web links *are not* permanent (even *permalinks*)

there is no general guarantee that a URL… which at one time points to a given object continues to do so
T. Berners-Lee et al. Uniform Resource Locators. RFC 1738.

URLs used in articles *decay*!

Analysis of *IEEE Computer* (Computer), and the *Communications of the ACM* (CACM): 1995-1999

- the *half-life* of a referenced URL *is approximately 4 years* from its publication date
  D. Spinellis. The Decay and Failures of URL References.
  Communications of the ACM, 46(1):71-77, January 2003.

Similar findings in Lawrence, S. et al. *Persistence of Web References in Scientific Research*, IEEE Computer, 34(2), pp. 26–31, 2001.

Scholar roster of broken links

An example from Astronomy

Domain	links (broken)	.html	.txt	.dat	.gz	.tar	.fits	tilde
cxc.harvard.edu	802 (110)	336 (70)	0	0	4 (2)	5 (4)	1	0
heasarc.gsfc.nasa.gov	640 (33)	423 (27)	1	0	0	0	0	0
www.stsci.edu	498 (61)	205 (29)	3	0	0	0	0	15 (10)
asc.harvard.edu	471 (152)	212 (99)	0	0	0	0	0	1 (1)
ssc.spitzer.caltech.edu	427 (194)	125 (76)	3 (3)	0	0	0	0	0
cfa-www.harvard.edu	352 (68)	277 (52)	1	0	0	0	0	54 (17)
archive.stsci.edu	308 (58)	57 (9)	2	1 (0)	0	0	0	0
www.ipac.caltech.edu	285 (14)	209 (12)	0	0	0	0	0	0
www.atnf.csiro.au	211 (21)	12 (6)	0	0	0	0	0	7 (5)
space.mit.edu	193 (10)	58 (5)	1	0	0	0	0	2 (1)
www.astro.psu.edu	186 (4)	103 (1)	1	10	1	1	0	2
www.eso.org	186 (58)	54 (22)	1 (1)	0	0	0	0	4 (1)
irsa.ipac.caltech.edu	163 (5)	38	0	0	1	0	0	0
www.sdss.org	156 (2)	106 (1)	0	0	0	0	0	0
hea-www.harvard.edu	125 (37)	42 (17)	1	0	0	1	0	26 (16)
physics.nist.gov	125 (3)	63 (2)	0	0	0	0	0	0
www.noao.edu	120 (3)	50 (2)	0	0	0	0	0	0
xmm.vilspa.esa.es	118 (35)	23 (19)	0	0	8 (1)	0	0	1 (1)
www.astro.princeton.edu	115 (31)	43 (14)	0	0	0	0	0	53 (12)
ad.usno.navy.mil	110 (27)	98 (22)	3 (3)	0	0	0	0	1 (1)

This table lists total number of links and broken links (HTTP status codes 3xx, 4xx, and 5xx) to top domains (domains with over 100 links) found within articles published in the four main astronomy journals between 1997 and 2008. The table also shows, for each domain, the portion of links to common filename extensions, as well as links that contain the tilde character.
doi:10.1371/journal.pone.0104798.t001

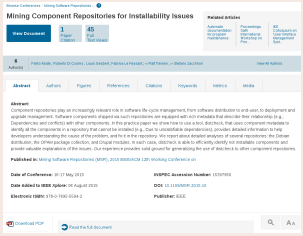
How Do Astronomers Share Data?
Pepe, Goodman, Muench, Crosas, Erdmann
dx.doi.org/10.1371/journal.pone.0104798

PLOS August 28, 2014

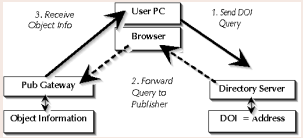
DOI limitations

Example: doi:10.1109/MSR.2015.10

- to find what 10.1109/MSR.2015.10 is, go to a *resolver* (e.g. doi.org)
- this returns `http://ieeexplore.ieee.org/document/7180064/`
- at this URL we find …



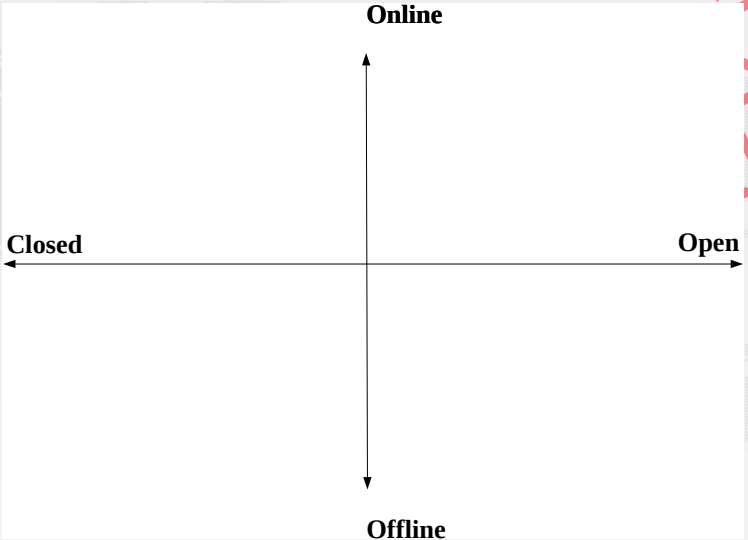
Architecture of the DOI infrastructure



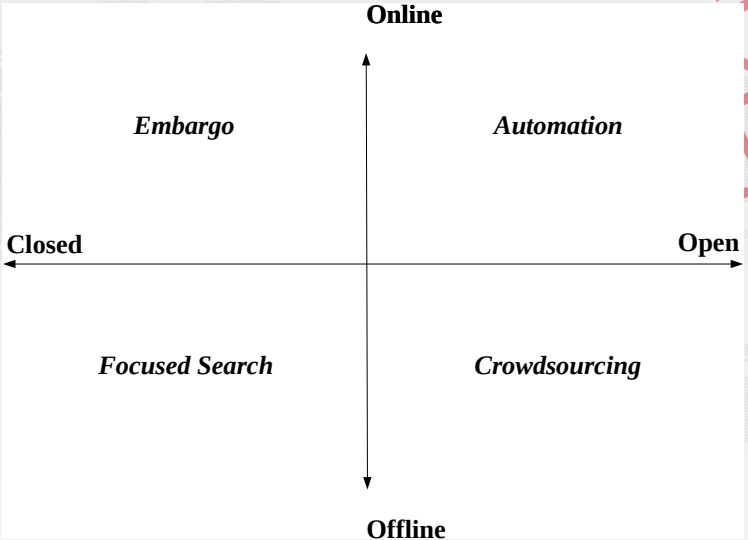
- DOI resolution *can change*
- content at URL *can change*
- no *intrinsic* way of noticing
- persistence based on *good will* of *multiple parties*

Outline

All the source code

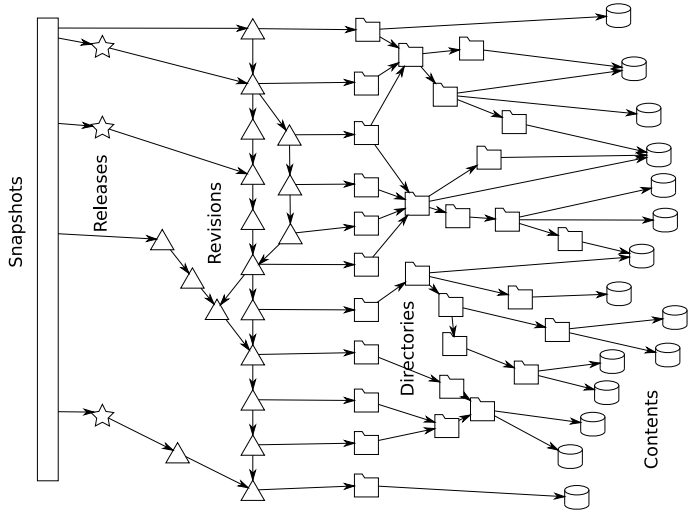


All the source code: strategy



Outline

The archive in pictures



Contents

A bird's eye view