Software Heritage
How to use it for Science and how to contribute

Roberto Di Cosmo
roberto@dicosmo.org

May 23rd, 2018
Outline

1. Software all around us
2. Source code for Science
3. The Software Heritage initiative
4. Using the Software Heritage archive for Science
5. Building for the long term
6. 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)

```c
float Q_rsqrt( float number )
{
  long i;
  float x2, y;
  const float threehalves = 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 * ( threehalves - ( x2 * y * y ) ); // 1st iteration
  // y = y * ( threehalves - ( x2 * y * y ) ); // 2nd iteration, this can be removed
  return y;
}
```

Net. queue in Linux (excerpt)

```c
/*
 * SFB uses two 8[i][j][k] : L x N arrays of bins (L levels, N bins per level)
 * This implementation uses L = 8 and N = 10
 * 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.”

Roberto Di Cosmo www.dicosmo.org

www.softwareheritage.org May 23rd, 2018 3 / 21
"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 Guidance Computer (~60,000 lines), 1969

Linux Kernel

... now in your pockets!

are we taking care of all this?
Software is spread all around
Software is fragile
Collberg’s report from the trenches

Analysis of 613 papers

- 8 ACM conferences: ASPLOS’12, CCS’12, OOPSLA’12, OSDI’12, PLDI’12, SIGMOD’12, SOSP’11, VLDB’12
- 5 journals: TACO’9, TISSEC’15, TOCS’30, TODS’37, TOPLAS’34

all very practical oriented

The basic question

can we get the code to build and run?
The result

This can be debated (see http://cs.brown.edu/~sk/Memos/Examining-Reproducibility/), but...

... that’s a whopping 81% of non reproducible works!
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


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


### An example from Astronomy

<table>
<thead>
<tr>
<th>Domain</th>
<th>links (broken)</th>
<th>.html</th>
<th>.txt</th>
<th>.dat</th>
<th>.gz</th>
<th>.tar</th>
<th>.fits</th>
<th>tilde</th>
</tr>
</thead>
<tbody>
<tr>
<td>cse.harvard.edu</td>
<td>802 (110)</td>
<td>336 (70)</td>
<td>0</td>
<td>0</td>
<td>4 (2)</td>
<td>5 (4)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>hessarc.gsfc.nasa.gov</td>
<td>640 (33)</td>
<td>421 (27)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.stsci.edu">www.stsci.edu</a></td>
<td>496 (61)</td>
<td>205 (29)</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15 (10)</td>
</tr>
<tr>
<td>asc.harvard.edu</td>
<td>471 (112)</td>
<td>212 (99)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (1)</td>
</tr>
<tr>
<td>noc.ipac.caltech.edu</td>
<td>427 (194)</td>
<td>135 (70)</td>
<td>3 (3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cfa-www.harvard.edu</td>
<td>352 (68)</td>
<td>277 (52)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>54 (17)</td>
</tr>
<tr>
<td>archive.stsci.edu</td>
<td>306 (58)</td>
<td>57 (9)</td>
<td>2</td>
<td>1 (8)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.ipac.caltech.edu">www.ipac.caltech.edu</a></td>
<td>285 (14)</td>
<td>209 (12)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.swift.edu">www.swift.edu</a></td>
<td>211 (21)</td>
<td>12 (6)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7 (5)</td>
</tr>
<tr>
<td>space.mit.edu</td>
<td>193 (10)</td>
<td>58 (5)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2 (1)</td>
</tr>
<tr>
<td><a href="http://www.astroucs.edu">www.astroucs.edu</a></td>
<td>186 (6)</td>
<td>103 (1)</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><a href="http://www.eso.org">www.eso.org</a></td>
<td>186 (58)</td>
<td>54 (22)</td>
<td>1 (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4 (1)</td>
</tr>
<tr>
<td>cfa-ipac.caltech.edu</td>
<td>163 (5)</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.sdss.org">www.sdss.org</a></td>
<td>156 (2)</td>
<td>106 (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hea-www.harvard.edu</td>
<td>125 (37)</td>
<td>42 (17)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>26 (16)</td>
</tr>
<tr>
<td>physics.nist.gov</td>
<td>125 (3)</td>
<td>63 (2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.nrao.edu">www.nrao.edu</a></td>
<td>120 (3)</td>
<td>50 (2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.ubpa.esa.es">www.ubpa.esa.es</a></td>
<td>116 (35)</td>
<td>23 (19)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (1)</td>
</tr>
<tr>
<td><a href="http://www.atmos.princeton.edu">www.atmos.princeton.edu</a></td>
<td>115 (31)</td>
<td>43 (14)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>53 (12)</td>
</tr>
<tr>
<td><a href="http://www.usno.navy.mil">www.usno.navy.mil</a></td>
<td>110 (27)</td>
<td>98 (22)</td>
<td>3 (3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (1)</td>
</tr>
</tbody>
</table>

This table lists total number of links and broken links (HTTP status codes 30x, 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.

dx.doi.org/10.1371/journal.pone.0104798

---

**How Do Astronomers Share Data?**

Pepe, Goodman, Muench, Crosas, Erdmann

dx.doi.org/10.1371/journal.pone.0104798

**PLOS August 28, 2014**

---
**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*
We are at a turning point

**Looking at the past**

- a lot of old software misplaced, lost, or behind barriers, but…
- most founding fathers are still here, and willing to share
- **urgent** to collect their knowledge

Only a few years left.

**Looking at the future**

- software development and use skyrockets: more programmers, and more code!
- **essential** to provide a *universal* platform for all the future software source code

Every year that goes by makes the problem worse.

**it is urgent to take action!**
Outline

1. Software all around us
2. Source code for Science
3. The Software Heritage initiative
4. Using the Software Heritage archive for Science
5. Building for the long term
6. Conclusion
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


Software Heritage

- **Technology**
  - transparency and FOSS
  - replicas all the way down

- **Content**
  - intrinsic identifiers
  - facts and provenance

- **Organization**
  - non-profit
  - multi-stakeholder
- full development history permanently archived
- origins: GitHub (automated), Debian (automated), Gitorious, Google Code, GNU
- ~200Tb raw contents, ~10Tb graph (7+Bn nodes, 60+Bn edges)
Reference archive for (scientific) software

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


If no time for a demo, let’s highlight some features…

Origin search

Directory browsing

Revisions as diffs
Deposit Scientific Software

Deposit software in HAL


Generic mechanism:
- SWORD based
- review process
- versioning

How to do it:
- today: deposit .zip file
- tomorrow:
  - provide SWH id and metadata
  - provide SWH id, metadata is extracted
  - ...

The way to go for publishing research software, spread the word!
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 evolutiona!

Coming soon to a platform near you!
Outline

1. Software all around us
2. Source code for Science
3. The Software Heritage initiative
4. Using the Software Heritage archive for Science
5. Building for the long term
6. Conclusion
Growing Support

Landmark Inria Unesco agreement, April 3rd, 2017

Sharing the vision

Contributing to the mission

> = 100K€/year
> = 50K€/year
> = 25K€/year
> = 10K€/year
You can help!

Connect Inria’s forges! see http://bit.ly/swhlisters

★★★ install FusionForge plugin, develop a lister
★★★ develop Gitlab plugin, develop lister (also for framagit!)

Contribute to the development see http://forge.softwareheritage.org

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

Spread the word!

• help research teams use the archive
• tell everybody about Software Heritage

Funding

• pester companies to become sponsors: sponsorship.softwareheritage.org
• give your own contribution: www.softwareheritage.org/donate
1. Software all around us
2. Source code for Science
3. The Software Heritage initiative
4. Using the Software Heritage archive for Science
5. Building for the long term
6. 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