Software is everywhere…

... and we are not taking care of it!

The Software Heritage initiative

Architecture

Open Science

Building for the long term

Conclusion
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)

```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 B[i][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 32 bit hash (provided per packet by rxhash or
 * external classifier) into 8 subshashes 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.”
So/f_tware is everywhere...

... and we are not taking care of it!

The Software Heritage initiative

Architecture

Open Science

Building for the long term

Conclusion
Software is spread all around
Software is fragile

damage
disaster
malicious
obsolete
tear
attack
media
aging
dependencies
deflection
storage
reference
corruption
corrupption
format
encryption
encryption
wear
dangling
Software lacks its own research infrastructure

Photo: ALMA(ESO/NAOJ/NRAO), R. Hills
1. Software is everywhere...
2. ... and we are not taking care of it!
3. The Software Heritage initiative
4. Architecture
5. Open Science
6. Building for the long term
7. Conclusion
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


Technology
- transparency and FOSS
- replicas all the way down

Content
- intrinsic identifiers
- facts and provenance

Organization
- non-profit
- mirror network

Software Heritage

Source files
- 4,536,067,027

Commits
- 1,024,675,748

Projects
- 83,801,775

Roberto Di Cosmo
www.dicosmo.org

(CC-BY 4.0) www.softwareheritage.org September 20th, 2018
1. Software is everywhere...
2. ... and we are not taking care of it!
3. The Software Heritage initiative
4. Architecture
5. Open Science
6. Building for the long term
7. Conclusion
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)
1. Software is everywhere...
2. ... and we are not taking care of it!
3. The Software Heritage initiative
4. Architecture
5. Open Science
6. Building for the long term
7. Conclusion
Research software: a long way to go!

ICSE (Zannier, Melrik, Maurer, 2006)
- complete absence of replication studies

- 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
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>.flts</th>
<th>tilde</th>
</tr>
</thead>
<tbody>
<tr>
<td>cse.harvard.edu</td>
<td>802 (110)</td>
<td>336</td>
<td>0</td>
<td>0</td>
<td>4 (2)</td>
<td>5 (4)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>fhexarc.gsfc.nasa.gov</td>
<td>640 (33)</td>
<td>423</td>
<td>27</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>490 (931)</td>
<td>205</td>
<td>290</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>asc.harvard.edu</td>
<td>471 (152)</td>
<td>212</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (1)</td>
</tr>
<tr>
<td>jpc.spitzer.caltech.edu</td>
<td>427 (194)</td>
<td>125</td>
<td>76</td>
<td>3</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</td>
<td>52</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>54 (17)</td>
</tr>
<tr>
<td>archive.stsci.edu</td>
<td>308 (58)</td>
<td>57</td>
<td>9</td>
<td>2</td>
<td>1</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</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.anst.csiro.au">www.anst.csiro.au</a></td>
<td>211 (21)</td>
<td>12</td>
<td>6</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>195 (10)</td>
<td>58</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2 (1)</td>
</tr>
<tr>
<td><a href="http://www.astropiu.edu">www.astropiu.edu</a></td>
<td>186 (4)</td>
<td>103</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>1</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</td>
<td>22</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4 (1)</td>
</tr>
<tr>
<td>isa.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.seds.org">www.seds.org</a></td>
<td>156 (2)</td>
<td>106</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tee-www.harvard.edu</td>
<td>125 (37)</td>
<td>42</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>physics.nist.gov</td>
<td>125 (3)</td>
<td>63</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.noea.edu">www.noea.edu</a></td>
<td>120 (3)</td>
<td>50</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>umm.ulsps.esa.es</td>
<td>116 (35)</td>
<td>23</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8 (1)</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.stanford.edu">www.stanford.edu</a></td>
<td>115 (31)</td>
<td>43</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>53 (12)</td>
</tr>
<tr>
<td>ad.usno.navy.mil</td>
<td>110 (27)</td>
<td>98</td>
<td>22</td>
<td>3</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 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

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

Many failed approaches, several reasons

**Use of inappropriate identifiers, see bit.ly/swhpidpaper**
- DOIs, URLs, PURLs

**Lack of a universal archive**
- disconnected silos
- sustainability issues

**Nota bene**

*citing software* and *referencing software* are different concepts
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

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!)
- 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

Roberto Di Cosmo  www.dicosmo.org  
(© CC-BY 4.0)  www.softwareheritage.org  September 20th, 2018 17 / 21
Outline

1. Software is everywhere…
2. … and we are not taking care of it!
3. The Software Heritage initiative
4. Architecture
5. Open Science
6. Building for the long term
7. Conclusion
Growing Support

Landmark Inria Unesco agreement, April 3rd, 2017

Contributing to the mission

Sharing the vision
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

1. Software is everywhere...
2. ... and we are not taking care of it!
3. The Software Heritage initiative
4. Architecture
5. Open Science
6. Building for the long term
7. Conclusion
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
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

8 Science of Software

9 Strategy

10 Under the hood
All the source code: strategy

- Embargo
- Automation
- Focused Search
- Crowdsourcing

Online
Offline
Open
Closed
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
The archive in pictures

Snapshots

Releases

Revisions

Directories

Contents
The archive in pictures
Revisions

<table>
<thead>
<tr>
<th>SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author: Nicolas Dandrimont <a href="mailto:nicolas@dandrimont.eu">nicolas@dandrimont.eu</a> (Thu Sep 1 14:26:13 2016)</td>
</tr>
<tr>
<td>Committer: Nicolas Dandrimont <a href="mailto:nicolas@dandrimont.eu">nicolas@dandrimont.eu</a> (Thu Sep 1 14:26:13 2016)</td>
</tr>
<tr>
<td>Subject: proveance.tasks: add the revision -&gt; origin cache task</td>
</tr>
<tr>
<td>Parent: fc3a8b59ca1df424d860f2c29ab07fee4dc35d10</td>
</tr>
</tbody>
</table>

swh/storage/provenance/tasks.py

| 77 |

tree 515f00d44e92c65322aaa9bf3fa097c00dd9c7d
parent fc3a8b59ca1df424d860f2c29ab07fee4dc35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

proveance.tasks: add the revision -> origin cache task

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6
The archive in pictures
The archive in pictures

Roberto Di Cosmo
www.dicosmo.org

September 20th, 2018

Releases

object c0c9f6b1e134f593e7567570a1761b1566e6eb1d
type commit
tag v0.0.51
tagger Nicolas Dandrimont <nicolas@dandrimont.eu> 1472042163 +0200

Release swh.storage v0.0.51

- Add new metadata column to origin_visit
- Update swh-add-directory script for updated API

---BEGIN PGP SIGNATURE-----

---END PGP SIGNATURE-----

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb
The archive in pictures
Snapshots

```
commit 0fffeb25770:190525eb3ce21691466c53a1d9158 refs/heads/atime
commit ba54a3a24e39f0a323a6c292ce4f6be63c6eb refs/heads/directory-listing-arrays
commit d66e68d8f922383ff550882b7fe1c085d72238d9c5 refs/heads/foot
commit c7f79f9ee9682f8949508f5a50819f67e468e08 refs/heads/master
commit 7e4a197c46e2e2049e7a5b1ea9eb8a44361b0fc2 refs/heads/tmp-directory-add
commit 642a206f37d0e858b8a58d47b53e0ed5b225e82e refs/heads/tmp/generic-releases

tag 28f843b1370cf768e96597979f04907c7571755 refs/tags/v0.1

tag 72a21991a304e39996b080670f0f7bee72aee2c2d refs/tags/v0.10

tag 3598ec8ac8eeb0b70e3b37695a2380b4a47a5cc refs/tags/v0.11

tag 33378427a403b5a59a6777b70d58f6674fb76556 refs/tags/v0.12

tag 067f46e52755b27c7f590311c2bfa836c3f3eb435d refs/tags/v0.13

tag 5a63255868b54b1854b40e7f70d2b1a1e7f3b0c refs/tags/v0.14

tag 586fbae680b4475a0b5f9396743c3e58a1c9c7f refs/tags/v0.15

tag 0c80b085f4798eb3f36317742b2a283f6e0e5b51c refs/tags/v0.16

tag a5424da4ea3f8be5d35e5f2b6a6e3c9b6a7c7d6 refs/tags/v0.17

tag 278a21f654cd12222e565594962e1e86ff6c4993d9 refs/tags/v0.18

tag 606f4794a4506e979cfcd24a980000ce8263ebef47c refs/tags/v0.19

tag 32b05a50f3ca323bae0d7f5a5d0a382b37b487 refs/tags/v0.2

tag 3147c3d35c31e464f2f88e1e98eb1237edbeff27c refs/tags/v0.20

tag 215ea093a141e862e9727e6b4e6073a87988 refs/tags/v0.21

tag 3f0b168c287ba642522a247a1ed3df9f5f5afa1d refs/tags/v0.22

tag 8cbee8da0d73c5d262789e468a16ac3c72aba4 refs/tags/v0.23
...
```

id: b464cad1b66fff266a37b46ea6e7a04b545e904b