1. Software is everywhere…
2. … and we are not taking care of it!
3. The Software Heritage initiative
4. Architecture
5. Open Science
6. Science of Software
7. Building for the long term
8. 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.”
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. Science of Software
7. Building for the long term
8. Conclusion
Software is spread all around
Software is fragile

- damage
- disaster
- malicious
- obsolete
- aging
- attack
- media
- reference
- deletion
- corruptions
- encryption
- format
- wear
- storage
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. Science of Software
7. Building for the long term
8. Conclusion
The Software Heritage Project

www.softwareheritage.org

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

1 Software is everywhere...
2 ... and we are not taking care of it!
3 The Software Heritage initiative
4 Architecture
5 Open Science
6 Science of Software
7 Building for the long term
8 Conclusion
### Research software: a long way to go!

<table>
<thead>
<tr>
<th>ICSE (Zannier, Melrik, Maurer, 2006)</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete absence of replication studies</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACM TOSEM 2001 to 2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>60% of all papers have tools: <strong>only 20% installable</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Collberg’s 2015 study</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://reproducibility.cs.arizona.edu/">http://reproducibility.cs.arizona.edu/</a></td>
</tr>
<tr>
<td>601 mainstream papers: 508 with tools, <strong>only 40% installable</strong></td>
</tr>
</tbody>
</table>

### 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

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

(13/18)
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. Science of Software
7. Building for the long term
8. Conclusion
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!
1. Software is everywhere…
2. … and we are not taking care of it!
3. The Software Heritage initiative
4. Architecture
5. Open Science
6. Science of Software
7. Building for the long term
8. Conclusion
Growing Support

Landmark Inria Unesco agreement, April 3rd, 2017

Roberto Di Cosmo www.dicosmo.org

Sharing the vision

Contributing to the mission

>= 100Ke/year
>= 50Ke/year
>= 25Ke/year
>= 10Ke/year

Microsoft
Intel
Societe Generale

GitHub
UQAM
FOSSID
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
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
1. Software is everywhere...
2. ... and we are not taking care of it!
3. The Software Heritage initiative
4. Architecture
5. Open Science
6. Science of Software
7. Building for the long term
8. 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
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

9. Replicability/traceability
10. Strategy
11. Under the hood
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


# Scholar roster of broken links

## 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>.fts</th>
<th>tilde</th>
</tr>
</thead>
<tbody>
<tr>
<td>ic.harvard.edu</td>
<td>802 (110)</td>
<td>336</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>hessarc.gsfc.nasa.gov</td>
<td>640 (33)</td>
<td>423</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>498 (91)</td>
<td>205</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>as.harvard.edu</td>
<td>471 (152)</td>
<td>212</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (1)</td>
</tr>
<tr>
<td>psc.spitzer.caltech.edu</td>
<td>427 (194)</td>
<td>125</td>
<td>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</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td>archive.stsci.edu</td>
<td>308 (58)</td>
<td>57</td>
<td>2</td>
<td>1</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</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.nmf.cais.au">www.nmf.cais.au</a></td>
<td>211 (21)</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7 (1)</td>
</tr>
<tr>
<td>space.mit.edu</td>
<td>195 (10)</td>
<td>58</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.astropu.es.us">www.astropu.es.us</a></td>
<td>186 (4)</td>
<td>103</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.eso.org">www.eso.org</a></td>
<td>186 (58)</td>
<td>54</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4 (1)</td>
</tr>
<tr>
<td>iss.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.nsf.org">www.nsf.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>bse-www.harvard.edu</td>
<td>125 (37)</td>
<td>42</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</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.noea.edu">www.noea.edu</a></td>
<td>120 (3)</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>umm.vulpes.es.es</td>
<td>118 (33)</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1 (1)</td>
</tr>
<tr>
<td><a href="http://www.astroph.princeton.edu">www.astroph.princeton.edu</a></td>
<td>115 (31)</td>
<td>43</td>
<td>0</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>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 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] [DOI]

---

# 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
All the source code

Roberto Di Cosmo www.dicosmo.org

(CC-BY 4.0) www.softwareheritage.org September 19th, 2018 5 / 8
All the source code: strategy

- Embargo
- Online
- Open
- Closed
- Automation
- Offline
- Open
- Crowdsourcing
- Offline
- Focused Search
- Closed
Outline

9 Replicability/traceability

10 Strategy

11 Under the hood