Making Software a first class citizen in the scholarly world

Archive - Reference - Describe - Cite (ARDC)

Roberto Di Cosmo
Inria and Université de Paris

SEFM 2020, Amsterdam
Outline

1. Software Source Code is Knowledge
2. Meet Software Heritage
3. Demo time!
4. The road ahead

Roberto Di Cosmo

(CC-BY 4.0) Research Software 17/09/2020 1 / 16
Software source code: a precious part of our heritage

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.”

Apollo 11 source code (excerpt)

<table>
<thead>
<tr>
<th>P63SPOT3</th>
<th>CA</th>
<th>BIT6</th>
<th># IS THE LR ANTENNA IN POSITION 1 YET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EXTEND</td>
<td>RAND</td>
<td>CHAN33</td>
</tr>
<tr>
<td></td>
<td>EXTEND</td>
<td>BZF</td>
<td>P63SPOT4</td>
</tr>
<tr>
<td></td>
<td>CAF</td>
<td>CODE500</td>
<td># ASTRONAUT: PLEASE CRANK THE</td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>BANKCALL</td>
<td># SILLY THING AROUND</td>
</tr>
<tr>
<td></td>
<td>CADR</td>
<td>GOPERF1</td>
<td># TERMINATE</td>
</tr>
<tr>
<td></td>
<td>TCF</td>
<td>GOTOP00M</td>
<td># PROCEED SEE IF HE’S LYING</td>
</tr>
<tr>
<td></td>
<td>TCF</td>
<td>P63SPOT3</td>
<td># ENTER INITIALIZE LANDING RADAR</td>
</tr>
</tbody>
</table>

Quake III source code (excerpt)

```c
float qr_sqrt(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;
}
```

Len Shustek, Computer History Museum

“Source code provides a view into the mind of the designer.”
Source code is special (software is not data)

Software evolves over time
- projects may last decades
- the development history is key to its understanding

Complexity
- millions of lines of code
- large web of dependencies
  - easy to break, difficult to maintain
  - research software a thin top layer
- sophisticated developer communities

Precious, endangered executable and human readable knowledge
key people passing away, platforms (GoogleCode, Gitorious, etc.) closing down …
no organised effort to catalog and archive it
So/f_tware: Source code: a pillar of Open Science

Three pillars of Open Science

C. Borgman, Paris, 2018

Sometimes, if you don't have the software, you don't have the data

A plurality of needs

Researcher
- archive and reference software used in articles
- find useful software
- get credit for developed software
- verify/reproduce/improve results

Laboratory/team
- track software contributions
- produce reports / web page

Research Organization
- know its software assets
- technology transfer
- impact metrics

N.B.: links are essential
Where we stand

**lack of reproducibility in SE and CS ..**

- no replication studies (Zannier et al., ICSE 2006)
- only 20% installable tools in TOSEM 2001 to 2006 (Ghezzi, ICSE 2009)
- 601 mainstream papers: 508 with tools, only 40% installable (Collberg, 2015)
  
  **main reasons**: source code (or the right version of it) cannot be found

**a recent awakening (~2010)**

- Policies: Artifact Evaluation (AEC), ACM Artifact Review and Badging, …
- Working groups: FORCE11, RDA, SPSO, …
- Generalist Repositories: FigShare, Zenodo, (but here software is just data)

but a lot is left to be done!
Archive
Research software artifacts must be properly archived
make sure we can retrieve them (reproducibility)

Reference
Research software artifacts must be properly referenced
make sure we can identify them (reproducibility)

Describe
Research software artifacts must be properly described
make it easy to discover and reuse them (visibility)

Cite/Credit
Research software artifacts must be properly cited (not the same as referenced!)
to give credit to authors (evaluation!)

We need an infrastructure designed for software source code: now we have one!
Outline

1. Software Source Code is Knowledge
2. Meet Software Heritage
3. Demo time!
4. The road ahead
Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

**Reference catalog**

*find* and *reference* all software source code

**Universal archive**

*preserve* all software source code

**Research infrastructure**

*enable analysis* of all software source code
An international, non profit initiative built for the long term

Sharing the vision

Donors, members, sponsors

Platinum sponsors

Gold sponsor

Silver sponsors

Bronze sponsors

And many more ...

www.softwareheritage.org/support/testimonials

(CC-BY 4.0) Research Software 17/09/2020 7 / 16
Experts call for greater recognition of software source code as heritage for sustainable development

6 November 2018

UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …

Their call is published on Feb 2019

It’s an important policy tool, already referenced and used …

yes, you can sign it!

The largest software archive, a shared infrastructure

Software Heritage

As of today the archive already contains and keeps safe for you the following amount of objects:

- **Source files**: 8,846,381,610
- **Commits**: 1,880,663,008
- **Projects**: 140,348,311
- **Directories**: 7,506,954,410
- **Authors**: 38,603,337
- **Releases**: 15,051,940
Addressing the four ARDC needs (see ICMS 2020 for details)

**Archive (8B+ files, 140M+ projects)**

- save.softwareheritage.org
- deposit.softwareheritage.org

**Reference (20 billion SWHIDs)**

Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs

```
prefix object_type
schema_version
object_id
```

Now supported in SPDX 2.2, Wikidata etc.

**Describe**

- **Intrinsic metadata** from source code
- Contributed the Codemeta generator

**Cite/Credit**

- Contributed *software citation* style
  biblatex-software, v 1.2-2 now on CTAN

Roberto Di Cosmo

(CC-BY 4.0) Research Software 17/09/2020 10 / 16
1. **Software Source Code is Knowledge**
2. **Meet Software Heritage**
3. **Demo time!**
4. **The road ahead**
A walkthrough

- Browse **the archive**
- Get and use SWHIDs (**full specification available online**)
  - Example use in a research article: compare Fig. 1 and conclusions
    - in the 2012 version
    - in the updated version using SWHIDs and Software Heritage
- Cite software **with the biblatex-software style** from CTAN
  - Example use in a research article: extensive use of SWHIDs in a replication experiment
- Trigger archival of your preferred software in a breeze
- Curated deposit in SWH via HAL, see for example: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, …
- Rescue landmark legacy software, see the **SWHAP process with UNESCO**
A word on the trust model for systems of identifiers

Two general classes of systems of identifiers

- **intrinsic** computed from the object (no registry required, fully decentralised)
  (e.g.: chemical notation, music notation, hashes, SWHIDs)

- **extrinsic** assigned by an authority (need a registry)
  (e.g.: passport number, DOI, ARK, RRID, etc.)

See the dedicated blog post for more details
1. Software Source Code is Knowledge
2. Meet Software Heritage
3. Demo time!
4. The road ahead
Adoption is coming …

HAL software curated deposit workflow

*Curated Archiving of Research Software Artifacts*

International Journal of Digital Curation, 2020

Reference archive for swmath.org

See code links, e.g.

SemiPar package

Image Processing On Line (IPOL)

- archives
- reference
- cite: see BibLaTeX example

JTCAM (Theor. Comp. and Appl. Mech)

- instructions for authors recommend archival in Software Heritage
- biblatex-software in journal \LaTeX{} class

Policy

now officially in the *French National Plan for Open Science*

Self archival guidelines

- online summary
- full ICMS 2020 paper

Roberto Di Cosmo
Breaking news, and a lesson to be learned

**Saving 250,000 endangered repositories...**
- summer 2019: BitBucket announce Mercurial VCS phase out
- fall 2019: Software Heritage teams up with Octobus (funded by NLNet, thanks!)
- july 2020: BitBucket erases 250,000 repositories
- august 2020: bitbucket-archive.softwareheritage.org is live

... preserving the web of knowledge

(Tweet is here)

**Bottomline**
*explicit deposit* is important, …
… and we must promote it…
… but will never be enough.

*(think also of all software dependencies!)*
Summing up: a revolutionary infrastructure designed for source code

Software Heritage

www.softwareheritage.org

**global source code archive**

- harvest *all* software, not just research software
- save code now to trigger archival on demand
- API for curated deposit

**Library of Alexandria of source code**

**universal intrinsic identifiers**

SWHIDs provide standard independent of version control systems

**uniform data model, full graph of development history**

enables large scale, big code research
Let’s all make research software a first class citizen!

- leverage Software Heritage in conferences, journals, AEC for archival and reference
- adopt and promote biblatex-software to cite software artifacts
- join the conversation on software citation and software evaluation criteria
- tackle the scientific problems: big code, classification, infrastructure, etc.

Thank you!

R. Di Cosmo
Archiving and Referencing Source Code with Software Heritage
ICMS 2020 (https://dx.doi.org/10.1007/978-3-030-52200-1_36)

R. Di Cosmo, M. Gruenpeter, S. Zacchiroli
Referencing Source Code Artifacts: a Separate Concern in Software Citation,
CiSE 2020 (10.1109/MCSE.2019.2963148) (hal-02446202)

P. Alliez, R. Di Cosmo, B. Guedj, A. Girault, M.-S. Hacid, A. Legrand and N. Rougier
Attributing and referencing (research) software: Best practices and outlook from Inria,
CiSE 2020 (10.1109/MCSE.2019.2949413) (hal-02135891)

J.F. Abramatic, R. Di Cosmo, S. Zacchiroli
Building the Universal Archive of Source Code, CACM, October 2018 (10.1145/3183558)
Appendix
Outline

5 Big code

6 SWHIDs by the example

7 News
Reference platform for *Big Code*

- unique observatory of all software development
- big data, machine learning paradise: classification, trends, coding patterns, code completion…

First datasets are available!

- full graph of software development (~20Bn nodes, ~200Bn edges) see Pietri, Spinellis, Zacchirolı, MSR 2019
  [https://dx.doi.org/10.1109/MSR.2019.00030](https://dx.doi.org/10.1109/MSR.2019.00030)
5 Big code

6 SWHIDs by the example

7 News
Software Heritage Identifiers (SWHID)

An emerging standard
- in Linux Foundation’s SPDX 2.2
- IANA registered, WikiData property P6138

Examples:
- Apollo 11 AGC excerpt,
- Quake III rsqrt

Roberto Di Cosmo
(CC-BY 4.0) Research Software 17/09/2020 3 / 7
A worked example
A worked example

Contents

sha1: 8624bcdae55baeef...
sha256: 8ceb4b9ee5aded...
sha1_git: 94a9ed024d385...
length: 35147
A worked example
A worked example

Directories

100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d6a34af6f52cf3cf6b0c2f7bd0648fbd255e77f AUTHORS
100644 blob 94a9ed824d3859793618152ea559a168bbcb85e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e8667751dfb95c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5eaf2eb4 Makefile
100644 blob 5c7e3a5bbdbb038682ba7793f440492ed9678bb3 Makefile.local
100644 blob 8617980629c2d4e6080404f99a749b605b3e67b README.db_testing
100644 blob 76b29f94c815e0869c414d38d78d7ce08ec514e README.dev
040000 tree ele10eece948af0b93adb0372afc89f12e92618a bin
040000 tree 83e56d0beaf7793c77a45a345c80fcb8af503013 debian
040000 tree a34c9c4ba213f0cedc67f9816348d2795557af5 docs
100644 blob f2a6d32c6135aa7287bbd76167b01df2ae4f1539 requirements.txt
100755 blob eee147c36caflbcb2d820da8dc026c5b68180bc setup.py
040000 tree 224bb4cf4c67fcad1d60bdf2d2b294e7e1abf3 sql
040000 tree 8631c9cd77be993168107ab5baf51f40c8300e swf
040000 tree 8fb985b56ba8ed692f1209b2773b474c61d66c1 utils

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
A worked example
A worked example

Revisions

<table>
<thead>
<tr>
<th>Details</th>
<th>Changes</th>
<th>Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author: Nicolas Dandrimont <a href="mailto:nicolas@dandrimont.eu">nicolas@dandrimont.eu</a> (Thu Sep 11 14:26:13 2016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Committer: Nicolas Dandrimont <a href="mailto:nicolas@dandrimont.eu">nicolas@dandrimont.eu</a> (Thu Sep 11 14:26:13 2016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject: provenance.tasks: add the revision -&gt; origin cache task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent: 963634dca6ba5dc37e3ee426ba091092c267f9f6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
 tre 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
parent fc3a8b59ca1df424d860f2c29ab07fee4dc35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

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

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6
```
A worked example

Releases

object cbc9f16b1e134f593e7567570a1761b156e6eb1d
  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 -----

IQiZBAABCAD8ABBqKvZTFNfxxaWYqGzOGRbmfRyaWlvbQZUXAGxqQ7A9WwMo2+
neqiwq/afq6509D4jKe+KXWAw6K7x325EEL71uy
ahpZ6p3j806p6acC1+Yn58Ftch3J2Ytroc3WXXWqrtxWNN5Mo6F7DDb8qphwh8AD5i2
ICBIU2jXUcD9I3eKFPwvzzXQ+hB0sMWY35Dr6l6W7Z74M6uIPGgLhuHP755yo
lEnWnvo7Y1Ym61h56B7s3iXRaqA-i=becq6ub7T7Qyjl+pUq5cbymq3h7vri/FL
qsz2mn5rykz9r8iGh1+jpv+g55WlnPo35TH0uojlVqPKzh5PQPqu4hjIZskca
kJ6kAWyU80Meb+xNKVej15R3+i+yWB63QP5a13Vw0OTh6FndALCMnEmQlkC5dKmt
cl+Max17l3/gEdhfnQPW7G65DwXkBfKfhgVLOq3nV3Gq0QT7n1pq5M02c06H9+xwZc
Gq/K1rdht4hxnOh6dwPZyye0U5XGeuduW03f4Z5z4Wm+5mMzdck7v5J5O7tn
RqT7Us5XUeXg9hgZDgpxh5YTnp1g3PC76USTKoaGe4A4Zm1k0gm6rw8CVRhvYo
rnhb5SBNMmzpyF6y75opUb1R70ptFRRUGKWEtRGhK9IhwxKWzGKzKz67fzU9
qul5wZ7mZQbZ739nonAL+HvPAlhysCMeqUgh82zf+EIVIVxk=
=K0xP
---- END PGP SIGNATURE -----

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb
A worked example
id: b464cad1b66ff266a37b46ea6e7a04b545e904b
News: archiving public code

https://code.etalab.gouv.fr
Paris Call on Software Source Code

“We call to support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive”

SWHAP: an important step forward

- detailed guidelines to curate landmark legacy source code and archive it on Software Heritage
- intense cooperation with Università di Pisa and UNESCO
- open to all, we’ll promote it worldwide

https://www.softwareheritage.org/swhap
Thomas Jefferson, February 18, 1791

…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.

Welcoming ENEA

- first institutional mirror
- increased resilience
- AI infrastructure for researchers
- stepping stone to an European joint effort