Software Heritage: infrastructure for Open Science

hands on demo

Roberto Di Cosmo

Director, Software Heritage
Inria and Université de Paris Cité

January 2023

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- *30+ years* of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- *20+ years* of Free and Open Source Software
- *10+ years* building and directing structures for the common good

1999 *DemoLinux* – first live GNU/Linux distro

2007 *Free Software Thematic Group*
150 members 40 projects 200Me

2008 *Mancoosi project* www.mancoosi.org

2010 *IRILL* www.irill.org

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

2021 *EOSC Task Force on Infrastructures for Software*, European Union

Software and source code (reminder)

"The source code for a work means the preferred form of the work for making modifications to it."

— GPL Licence

Hello World

Program (excerpt of binary)

```
4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
4004f9: 90
4004fa: 5d
4004fb: c3
```

Program (source code)

```
/* Hello World program */

#include<stdio.h>

void main()
{
    printf("Hello World");
}
```

The *knowledge* is in the *source code*

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)

```
P63SPOT3        CA      BIT6            # IS THE LR ANTENNA IN POSITION 1 YET
                EXTEND
                RAND    CHAN33
                EXTEND
                BZF     P63SPOT4        # BRANCH IF ANTENNA ALREADY IN POSITION 1

                CAF     CODE500         # ASTRONAUT:    PLEASE CRANK THE
                TC      BANKCALL        #               SILLY THING AROUND
                CADR    GOPERF1
                TCF     GOTOPOOH        # TERMINATE
                TCF     P63SPOT3        # PROCEED       SEE IF HE'S LYING

P63SPOT4        TC      BANKCALL        # ENTER         INITIALIZE LANDING RADAR
                CADR    SETPOS1

                TC      POSTJUMP        # OFF TO SEE THE WIZARD ...
                CADR    BURNBABY
```

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;
}
```

Len Shustek, Computer History Museum (2006)

"Source code provides a view into the mind of the designer."

Source code in Open Science: a plurality of needs to address

Researchers

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

Laboratories/teams

- track software contributions
- produce reports
- maintain web page

Research Organizations and/or Funders

know its software assets

- technology transfer
- impact metrics
- funding strategy
- career evaluation

Archive, Reference, Describe, Cite and Credit

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*!)

Outline

Some old popular approaches

A - Since the ~~1970's~~ 1990's

.zip or .tar file on:

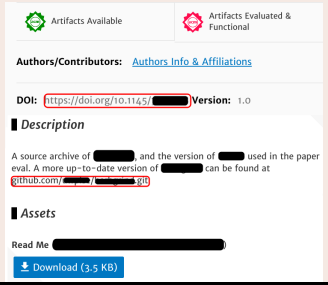
- ~~ftp server~~ (e.g. gnu)
- web page (e.g. myself)
- document archive (+ DOI sample)

B - Since the 2000's

Rely on *software forges*

- institutional/project (e.g. myself)
- free commercial ones: BitBucket, GitHub, GitLab, … (e.g. myself)

C: a mix of the two



Can get no satisfaction…

A *Poor user experience*

B *No preservation guarantee*

C Can do *so much* better

Forges are *not* archives!

2015: the first big bad news

Google Code and Gitorious.org shutdown: ~1M endangered repositories

- broken links in the web of knowledge (my papers too)

Big bad news keep coming in

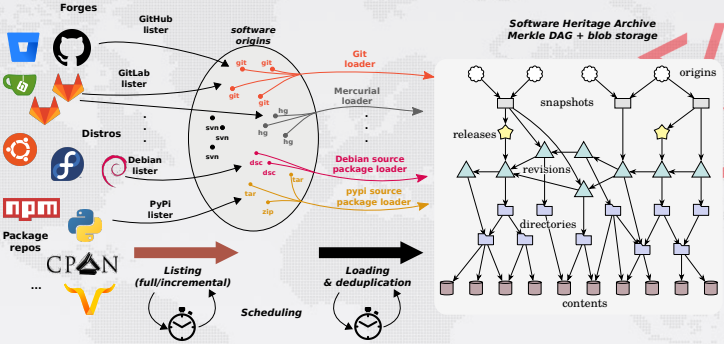
- summer 2019: BitBucket announces Mercurial VCS sunset
- july 2020: BitBucket erases *250.000+* repositories (including research software)
- summer 2022: GitLab.com considers erasing all projects that are inactive for a year

In Academia too!

- 2021: Inria's old gforge is unplugged… breaks the Opam build chain for OCaml

We need a universal archive of software source code: now we have one!

Software Heritage is *radically different*



Global development history permanently archived in a uniform data model

- over 13 billion unique source files from over 200 million software projects
- ~1PB (uncompressed) blobs, ~25 B nodes, ~350 B edges

Software Heritage is *radically different*, cont'd

A quick tour as a user

- designed for source code: Browse (e.g. Apollo 11 excerpt, see also Apollo 11 blog post) like on a developer platform, not a document archive!
- reference source code: all granularities, using SWHIDs (full specification available online)
  - compare Fig. 1 and conclusions in the 2012 version and the updated version
  - SWHID in a replication experiment
  - guidelines and a full article
  - SWHIDs *guarantee integrity* like in *blockchains*
    demo if time left:
    1. download a version of a project for a given SWHID
    2. compute locally the SWHID with swh-identify
    3. check that the computed id match the given one

Software Heritage is *radically different*, cont'd

Getting software archived

- automated harvesting: over 200 million software origins, your researchers' work may already be there (actually, here)!
- universal archive: *all* source code from *all* platforms (BitBucket, GitHub, GitLab, your own forge, etc.)
  - trigger archival of *any code* in one click with the updateswh browser extension
  - use webhooks to automatically archive *your code* (a GitHub action is available too)
  - journals, libraries, open access portals may *deposit sourcecode and metadata*
    - Example article from IPOL
    - Example article from eLife

A look at some adoption indicators

From Melissa Harrison's OSEC 2022 talk

What are they "referencing"?

source	n	percentage
Not available	2868	46.22
GitHub	1151	18.55
software heritage	387	6.24
zenodo	142	2.29
r package	70	1.13
cran	56	0.90
r package version	54	0.87
gitlab	35	0.56

- 6205 "software" references identified
- Top 8 listed, then long tail of 1055 other sites – 932 are unique "source"

Upcoming on replicabilitystamp.org (please do not spoil :-))

HAL+SWH in the Open Science software booklet

Funding agencies recommendations ANR 2023 guidelines (p. 17)

Enfin, conformément au 2ème Plan national pour la science ouverte, L'ANR recommande que les logiciels développés durant le projet soient mis à disposition sous une licence libre[30] et que les codes sources soient stockés dans l'archive Software Heritage[31] en indiquant la référence au financement ANR.

Outline

Describe, cite, credit

Software metadata: codemeta.json

- example from Parmap, created using the Codemeta generator

Integration with the HAL national french open access archive

- Curated deposit: metadata quality due to moderation
  - all pieces of the puzzle together: one researcher does all the steps (Parmap)
- export of citation information for biblatex-software
- examples: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, ...
- generation of reports, cv, web pages: for Inria, for CNRS, for CNES, for LIRMM or for Rémi Gribonval using HalTools

Bottomline

Software Heritage + a *curated* metadata repository allows to address all needs …

- *researcher, engineer*: archival, reference, credit, CV etc. *with a little effort from them*
- *labs, organizations*: track and report software production in a simple way
- *technology transfer offices*: view the software production
- *national level*: a *curated* catalog of the software production

Outline

A very large telescope of software source code

Where does this source code come from?

- Licence compliance for tech transfer
- Plagiarism detection for source code

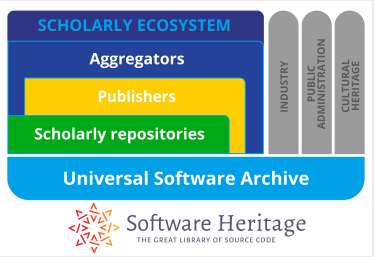
Software Heritage knows

- demo if time left
  - pick a random file in one of Roberto's projects
  - use swh-graph to find all the origins in SWH that contain it

Outline

Software Heritage: key infrastructure for Open Science

EOSC SIRS report: Software Source Code and Open Science, 2020



Connect scholarly ecosystem with the whole software ecosystem

See e.g. the French public administration open source catalog

Ongoing work: FAIRCORE4EOSC

A full workpackage:

- connectors with InvenioRDM, episcience, Dagstuhl, swMath, etc.
- Software Heritage mirror for the EOSC
- standardisation of CodeMeta and SWHID

Software Heritage: a shared open infrastructure

Software Heritage offers

- archival of all public source code
- reference of all public source code
- sharing cost with other partners
- standards based approach

Software Heritage is

- vendor neutral
- open source
- worldwide, long term
- born and based in the EU

Getting onboard: join the Deposit Interest Group

Four levels:

- Strategic
- Core
- Solutions
- Basic

- Flat membership fee
- Participation in advisory boards (next is February 7 2023)
- Access to working groups

Questions?

References (see https://www.softwareheritage.org/publications)

EOSC SIRS Task Force, *Scholarly Infrastructures for Research Software*
2020, European Commission, (10.2777/28598)

R. Di Cosmo, *Archiving and Referencing Source Code with Software Heritage*
ICMS 2020 (10.1007/978-3-030-52200-1_36)

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

Help Software Heritage grow and better serve you all … as well as all of humankind

← Learn more, become a member →

Thank you!