Archiving and Referencing all the software source code

Roberto Di Cosmo
Director, Software Heritage

ICMS 2020

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Software Source Code is knowledge
2. Software Heritage
3. Demo time!
4. Conclusion

Software source code: *human readable* and *executable knowledge*

Harold Abelson, Structure and Interpretation of Computer Programs (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)

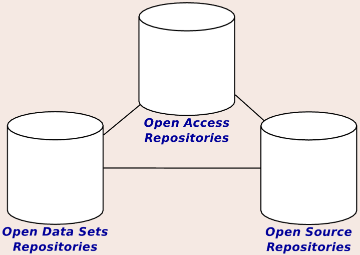
```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    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.”

Software Source code: pillar of Open Science

Three pillars of Open Science



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

What is at stake

in increasing order of difficulty

Archival

Research software artifacts must be properly archived

make sure we can *retrieve* them (*reproducibility*)

Identification

Research software artifacts must be properly referenced

make sure we can *identify* them (*reproducibility*)

Metadata

Research software artifacts must be properly described

make it easy to *discover* them (*visibility*)

Citation

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 it!

Outline

1. Software Source Code is knowledge
2. Software Heritage
3. Demo time!
4. Conclusion

Software Heritage, in a nutshell

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

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

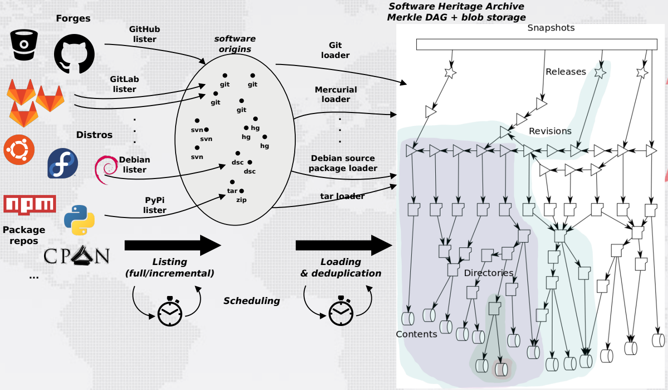
Universal archive

preserve all software source code

Research infrastructure

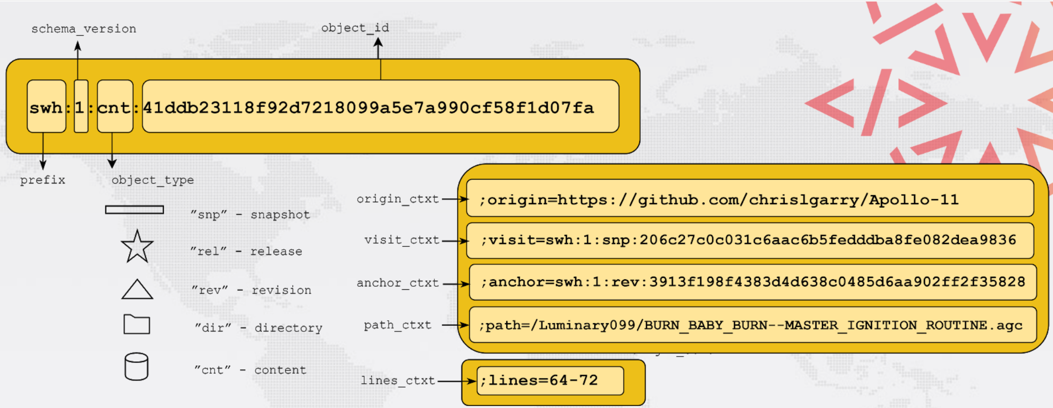
enable analysis of all software source code

Automation, and storage



- full development history permanently archived!
- over 8 billions unique source files from 130+ million origins

Software Heritage Identifiers (SWHID) link to full docs



schema_version

object_id

`swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa`

prefix

object_type

"snp" - snapshot

"rel" - release

"rev" - revision

"dir" - directory

"cnt" - content

origin_ctxt → `;origin=https://github.com/chrislgarry/Apollo-11`

visit_ctxt → `;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836`

anchor_ctxt → `;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828`

path_ctxt → `;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc`

lines_ctxt → `;lines=64-72`

An emerging standard

- in Linux Foundation's SPDX 2.2
- IANA registered, WikiData property P6138

Examples:

- Apollo 11 AGC excerpt,
- Quake III rsqrt

Software Heritage for research

Archive

- a *universal* archive: collects *all* software, not only academic software
- *harvests* source code worldwide *(8B+ files from 130M+ projects in July 2020)*
- your software may be there already... if not, please *save its code now*!

Reference

- SWHID: *intrinsic*, *decentralised*, *cryptographically strong* identifiers
- enhance articles with *source code references* for reproducibility

Cite

- biblatex-software : a dedicated bibliographic style for software!

Detailed guidelines in the paper and online!

Outline

1. Software Source Code is knowledge

2. Software Heritage

3. Demo time!

4. Conclusion

A walkthrough

- Browse the archive
- Get and use SWHIDs (full specification available online)
- cite software with the biblatex-software style from CTAN
- Example use in a research article: compare Fig. 1 and conclusions
  - in the 2012 version
  - in the updated version using SWHIDs and Software Heritage
- 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

Outline

1. Software Source Code is knowledge
2. Software Heritage
3. Demo time!
4. Conclusion

An international, non profit initiative built for the long term

Sharing the vision

UNESCO
United Nations
Educational, Scientific and
Cultural Organization

IEEE Software · GitLab · Engineering · eclipse · OW2 · fsfe · nlnet foundation · SIGSOFT · Liberté • Égalité • Fraternité République Française · INAPI · The Linux Foundation · Adullact · Informatics Europe · Computer History Museum · software freedom conservancy · Software Freedom Law Center · AdaCore The GNAT Pro Company · gandi.net · TDF · Free Software Foundation · creative commons · SIF Société Informatique de France · open source initiative · openinventionnetwork

And many more ...
www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Platinum sponsors

Microsoft · intel · Societe Generale · HUAWEI

Gold sponsor

Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation · openinventionnetwork · Université de Paris

Silver sponsors

CAST Software Intelligence for Digital Leaders · Liberté • Égalité • Fraternité République Française · GitHub · Google · Università di Pisa · vmware

Bronze sponsors

DANS Data Archiving and Networked Services · FOSSID · NOKIA Bell Labs · Alma Mater Studiorum A.D. 1088 · UQÀM Université du Québec à Montréal

The way forward

Software Heritage

- *universal* archive of source code
- *intrinsic* identifiers (SWHIDS)
- *open*, *non profit*, long term
- *infrastructure* for Open Science

You can help improve science!

- *use* SWH (see swmath.org and ipol.im)
- *save* relevant source code
- *contribute* to SWH: *it is open source*
- spread the word

📄 Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchiroli
Building the Universal Archive of Source Code, CACM, October 2018 (10.1145/3183558)

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

📄 Pierre Alliez, Roberto Di Cosmo, Benjamin Guedj, Alain Girault, Mohand-Said Hacid, Arnaud Legrand and Nicolas Rougier
Attributing and referencing (research) software: Best practices and outlook from Inria,
CiSE 2020 (10.1109/MCSE.2019.2949413) (hal-02135891)