Archiving and Referencing all the source code

working together to make software count

Roberto Di Cosmo
Director, Software Heritage

July 10th, 2020

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

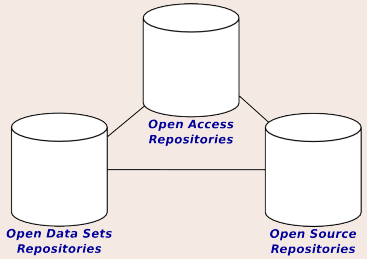
Outline

Software Source code: pillar of Open Science

Harold Abelson, Structure and Interpretation of Computer Programs (1985)

"Programs must be written for people to read, and only incidentally for machines to execute."

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

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

An international, non profit initiative built for the long term

Sharing the vision

United Nations
Educational, Scientific and
Cultural Organization

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

Donors, members, sponsors

INVENTEURS DU MONDE NUMÉRIQUE

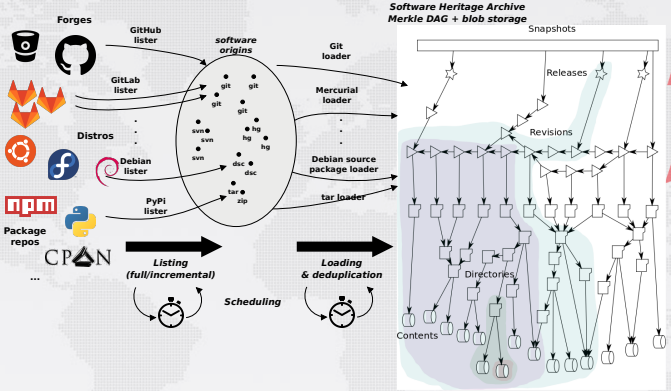
Platinum sponsors

Gold sponsor

Silver sponsors

Bronze sponsors

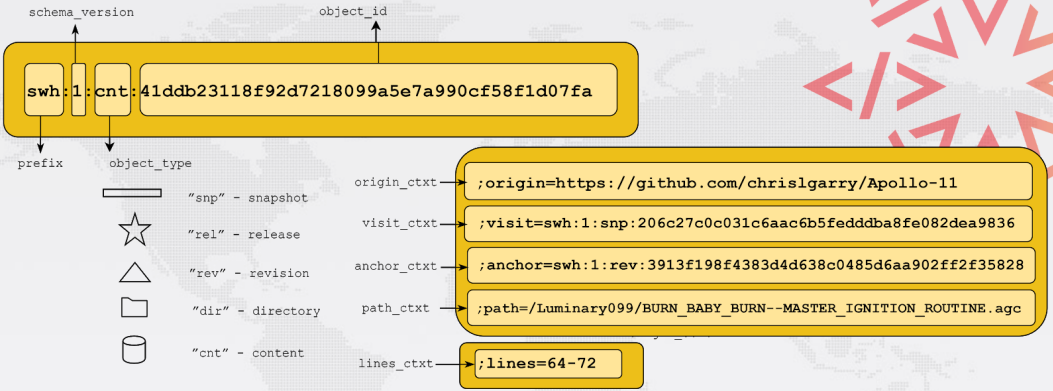
Automation, and storage



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

Software Heritage Identifiers (SWHID)

link to full docs



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
  - 8+ billion files from 130+ million projects in July 2020!
- your software is likely there already… and you can *save code now*

Reference

- SWHID: *intrinsic*, *decentralised*, *cryptographically strong* identifiers
  - for *over 20 billion artifacts*
  - IANA registered, adopted in industry and WikiData
- enhance articles with *source code references* for reproducibility

Detailed guidelines for researchers are available online!

Outline

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
- 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, ...
- cite software using the biblatex-software style
- rescue landmark legacy software, see the SWHAP process with UNESCO

Outline

The way forward

Software Heritage

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

Your help is needed!

adopt	*use* SWH in your work
save	relevant source code
contribute	SWH is open source
advocate	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)