Referencing (all) publicly available software source code

... with Software Heritage!

Stefano Zacchiroli

Université de Paris & Inria – zack@upsilon.cc, @zacchiro

9 September 2020

Workshop on Open Citations and Open Scholarly Metadata 2020
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*
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

Referencing (all) publicly available software source code
Software Heritage in a nutshell

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
The Software Heritage Archive

- Source files: 8,152,137,176
- Commits: 1,744,034,936
- Projects: 123,781,438

The richest public source code archive, and growing daily!
The Software Heritage Archive

- ~400 TB (uncompressed) blobs, ~20 B nodes, ~300 B edges
- The *richest* public source code archive, … and growing daily!
Saving and referencing research software

1. Prepare your public repository
   README, AUTHORS, & LICENSE files + metadata (e.g., CodeMeta)

2. Save your code
   https://save.softwareheritage.org

3. Reference your work
   full repository, specific version, or code fragment → using SWHIDs! (next slide)

Learn more

- *Saving and referencing research software in Software Heritage*
  on the Software Heritage blog, August 2019

Software Heritage Identifiers (SWHIDs)

- **IANA registered “swh:” URI prefix**
- **Wikidata property P6138**
- **Examples**
  - Apollo 11 AGC excerpt,
  - /Quake III rsqrt

Referencing (all) publicly available software source code
Software Heritage Identifiers (SWHIDs)

swlh:ctnt:41dbb23118f92d7218099a5e7a990cf58f1d07fa

prefix  object_type

- "snp" - snapshot
- "rel" - release
- "rev" - revision
- "dir" - directory
- "cnt" - content

IANA registered "swh:" URI prefix

Wikidata property P6138

Examples

Apollo 11 AGC excerpt,
/Q_uake III rsqrt

Stefano Zacchirolı

Referencing (all) publicly available software source code
Software Heritage Identifiers (SWHIDs)

Prefix: swh
Object Type: cnt
Object ID: 41db23118f92d7218099a5e7a990cf58f1d07fa

Examples:
- Apollo 11 AGC excerpt
- Quake III rsqrt

Stefano Zacchirolí

Referencing (all) publicly available software source code

WOOC 2020
Software Heritage Identifiers (SWHIDs)

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

**Standardization**
- Linux Foundation SPDX 2.2
- IANA registered "swh:" URI prefix
- Wikidata property P6138

```
swh:1:cnt:41dcb23118f92d7218099a5e7a990cf58f1d07fa
```

```
origin_ctx = ;origin=https://github.com/chrislgarry/Apollo-11
visit_ctx = ;visit=swh1:snp:206c27c031c6a6e6b5feddaba8fe082dea9836
anchor_ctx = ;anchor=swh1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828
path_ctx = ;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc
lines_ctx = ;lines=64-72
```
Software Heritage Identifiers (SWHIDs)

- Linux Foundation SPDX 2.2
- IANA registered "swh:" URI prefix
- Wikidata property P6138

Examples
- Apollo 11 AGC excerpt,
- Quake III rsqrt

Standardization
- Linux Foundation SPDX 2.2
- IANA registered "swh:" URI prefix
- Wikidata property P6138
Referencing software with SWHIDs

- **citing v. referencing** software are separate concerns in scholarly works
- **referencing** is an often neglected need, but a particularly important one in the context of scientific reproducibility
- **SWHID**: an identifier scheme to address source code referencing needs

Citing software with biblatex-software

- **biblatex-software**: a BibTeX extension to support citing software
- **citable artifacts**: software, software versions, software modules, code fragments
- support SWHID (where appropriate) to *reference* underling artifacts

Learn more

- *Citing software with style*, Software Heritage blog, May 2020
- *CTAN package documentation*
Wrapping up

- Software Heritage is the largest archive of public software source code. It supports scholars in archiving and referencing source code relevant to their work.

- Referencing and citing software are separate concerns in scholarly workflows.

- SWHID identifiers are an adopted standard to reference source code artifacts.

- biblatex-software allow to cite software artifacts and integrates well with SWHIDs.

Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchiroli
Building the Universal Archive of Source Code
Communication of the ACM, October 2018

Roberto Di Cosmo, Morane Gruenpeter, Stefano Zacchiroli
Referencing Source Code Artifacts: a Separate Concern in Software Citation
Computing in Science & Engineering, 2020, ISSN: 1521-9615

Roberto Di Cosmo
Archiving and Referencing Source Code with Software Heritage
International Congress on Mathematical Software (ICMS), 2020

Stefano Zacchiroli / zack@upsilon.cc / @zacchiro / @zacchiro@mastodon.xyz