

Referencing (all) publicly available software source code ... with Software Heritage !

Stefano Zacchiroli

Université de Paris & Inria – zack@upsilon.cc, [@zacchiro](https://twitter.com/zacchiro)

9 September 2020

Workshop on Open Citations
and Open Scholarly Metadata 2020



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE



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



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



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



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



GitHub



GitLab

Bitbucket

Google code



GITORIOUS

Framagit

HAL
archives-ouvertes.fr

debian

npm



GNU

Inria
inventeurs du monde numérique

python
Package Index


















- ~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

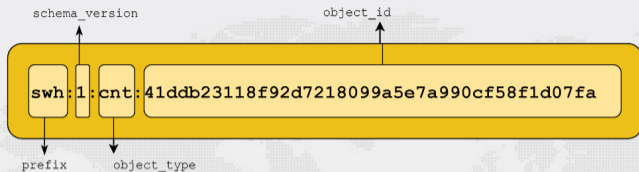
→ using SWHIDs ! (next slide)

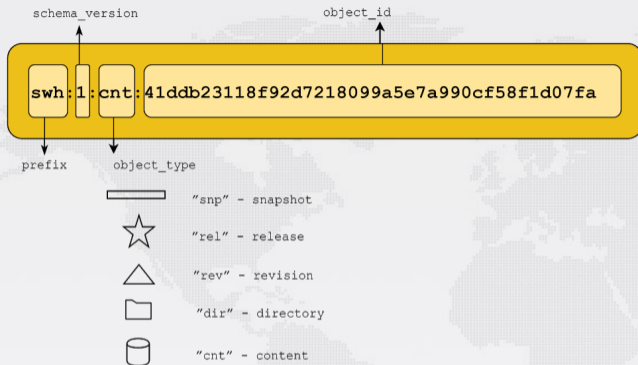
Learn more

- *Saving and referencing research software in Software Heritage* on the Software Heritage blog, August 2019
- *How to use Software Heritage for archiving and referencing your source code: guidelines and walkthrough* <https://annex.softwareheritage.org/public/guidelines/archive-research-software.pdf>

Software Heritage Identifiers (SWHIDs)

(link to full spec)









Standardization

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



Standardization

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

Examples

- Apollo 11 AGC excerpt,
- Quake III rsqrt

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

(sample .bib)

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