

Archive, reference, describe and cite software source code: a pathway to reproducibility

Morane Gruenpeter,
Software Heritage, Inria



Software Heritage

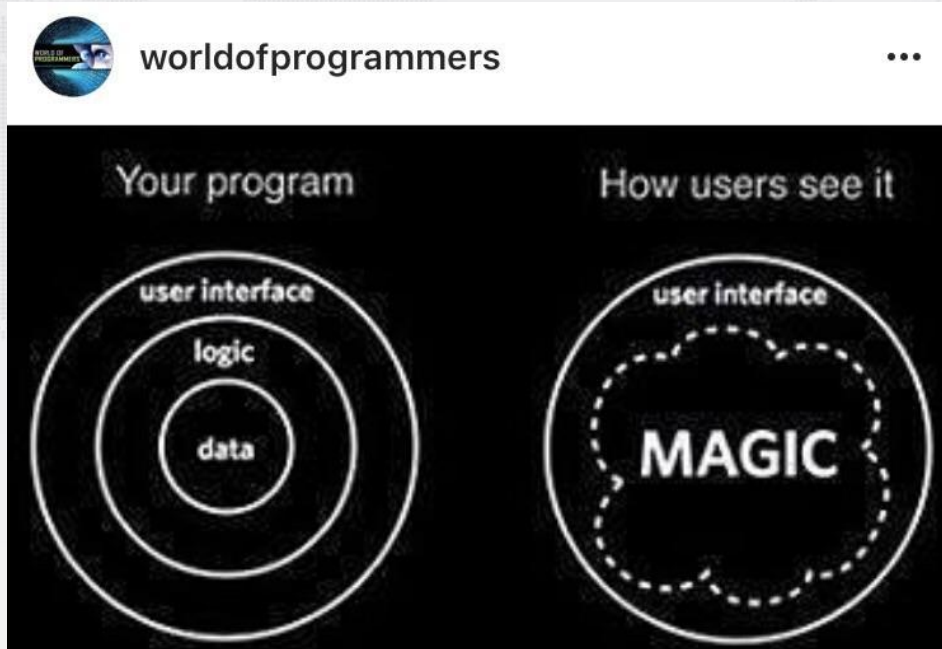
THE GREAT LIBRARY OF SOURCE CODE



Outline

- ★ **Preserving source code**
- ★ Software Heritage the universal source code archive
- ★ Good practices for software curation
 - Archive, reference, describe and cite
- ★ Conclusion

The Knowledge is in the source code



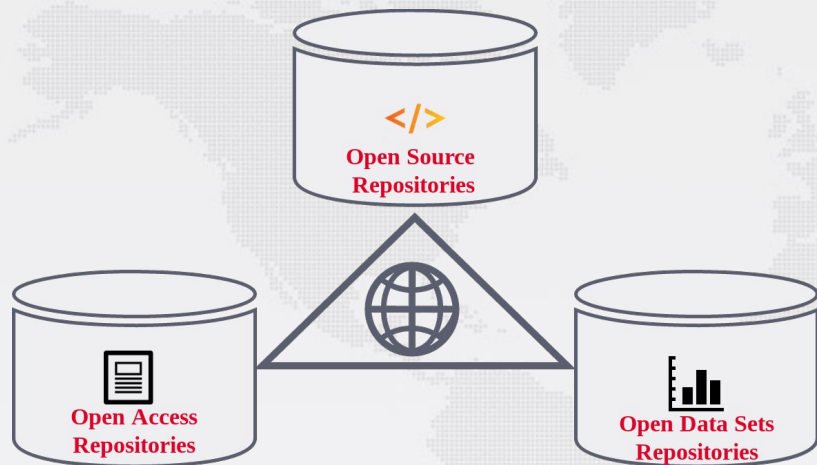
https://www.reddit.com/r/ProgrammerHumor/comments/70fump/programming_is_magic/



A precious part of our scientific heritage

Software source code has multiple facets in research, it can be seen as:

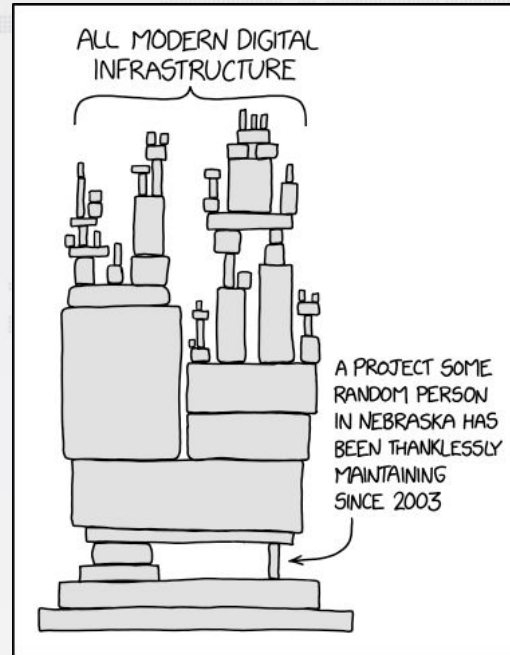
- a **tool**
- a research **outcome** or result
- **the object** of research



Three pillars of Open Science



In science, reproducibility requires long-term access to source code



<https://xkcd.com/2347/>

Forges are not archives!



Gabriel Altay
@gabrielaltay

Just realized [@Bitbucket](#) disabled all mercurial repositories when the [@asclnet](#) informed me that a link associated with an old paper of mine was down. Thought all was lost, but someone archived all the repos! very classy move by [@octobus_net](#) and [@SWHeritage](#).

[Traduire le Tweet](#)

1:48 AM · 31 août 2020 · Twitter Web App

August 2022 : [GitLab rumors](#)
about deleting inactive
projects from free accounts

Sunsetting Mercurial support in Bitbucket

April 21, 2020 | 3 min read



Denise Chan

[Update Aug 26, 2020] All hg repos have now been disabled and cannot be accessed.

[Update July 1, 2020] Today, mercurial repositories, snippets, and wikis will turn to read-only mode. After July 8th, 2020 they will no longer be accessible.

Source: [BitBucket blog](#)

250,000 repos



Software Heritage and Guix

Challenge:

How to ensure that a system can still be built even when the original source code becomes unavailable?

Solution:

Fall back on the universal source code archive

More on *Software Heritage to the rescue of reproducible science*

- Antoine R. Dumont, Valentin Lorentz (Software Heritage)





Outline

- ★ Preserving source code
- ★ Software Heritage the universal source code archive
- ★ Good practices for software curation
 - Archive, reference, describe and cite
- ★ Conclusion

The universal source code archive



Inside of Library of Alexandria-artistic rendition (Image: [21h007](#))



An international, non profit initiative



Diamond sponsor



Platinum sponsors



Gold sponsors



Silver sponsors



Bronze sponsors





The online archive - a must see

Full width

Home Development Documentation **Donate**

Operational login

We are hiring ! See our job offers.

 Software Heritage Archive

Features

 Search

 Downloads

 Save code now

 Add forge now

 Help

Welcome to the Software Heritage archive

Search

Enter a SWHID to resolve or string pattern(s) to search for in origin urls

☒ only show origins visited at least once

☒ filter out origins with no archived content

☐ search in metadata (instead of URL)

visit type: any

... or check our  Web API

Overview

The long term goal of the Software Heritage initiative is to **collect** all publicly available software in source code form together with its development history, replicate it massively to ensure its **preservation**, and **share** it with everyone who needs it. The Software Heritage archive is growing over time as we crawl new source code from software projects and development forges.

Content

A significant amount of source code has already been ingested in the Software Heritage archive. It notably includes the following software origins.

Regular crawling

These software origins get continuously discovered and archived using the **listers** implemented by Software Heritage.

Inside of Library of Alexandria-artistic archive.softwareheritage.org



Outline

- ★ Preserving source code
- ★ Software Heritage the universal source code archive
- ★ **Good practices for software curation**
 - Archive, reference, describe and cite
- ★ Conclusion



What is at stake?

Archive

- Research software artifacts must be properly archived
- make sure we can retrieve them (reproducibility)

Describe

- Research software artifacts must be properly described
- make it easy to discover them (visibility)

Reference

- Research software artifacts must be properly referenced
- make sure we can identify them (reproducibility)

Cite (for credit)

- Research software artifacts must be properly cited (not the same as referenced!)
- to give credit to authors (evaluation!)

Archive



HAL
open science

CCSD



Software Heritage



Submit code to HAL

Inria

- A scholarly repository
- An archive
- Curated metadata
- Transferring code to SWH

Save code now to SWH

- Easy (only submit repo's url)
- All dev history archived
- different vcs are accepted
- PID to reference specific pieces of code (even algorithms)

Classic deposit

.zip, .tar.gz

SWHID deposit

COMING SOON

Save code now

git, svn, hg, bzz

SWHID deposit

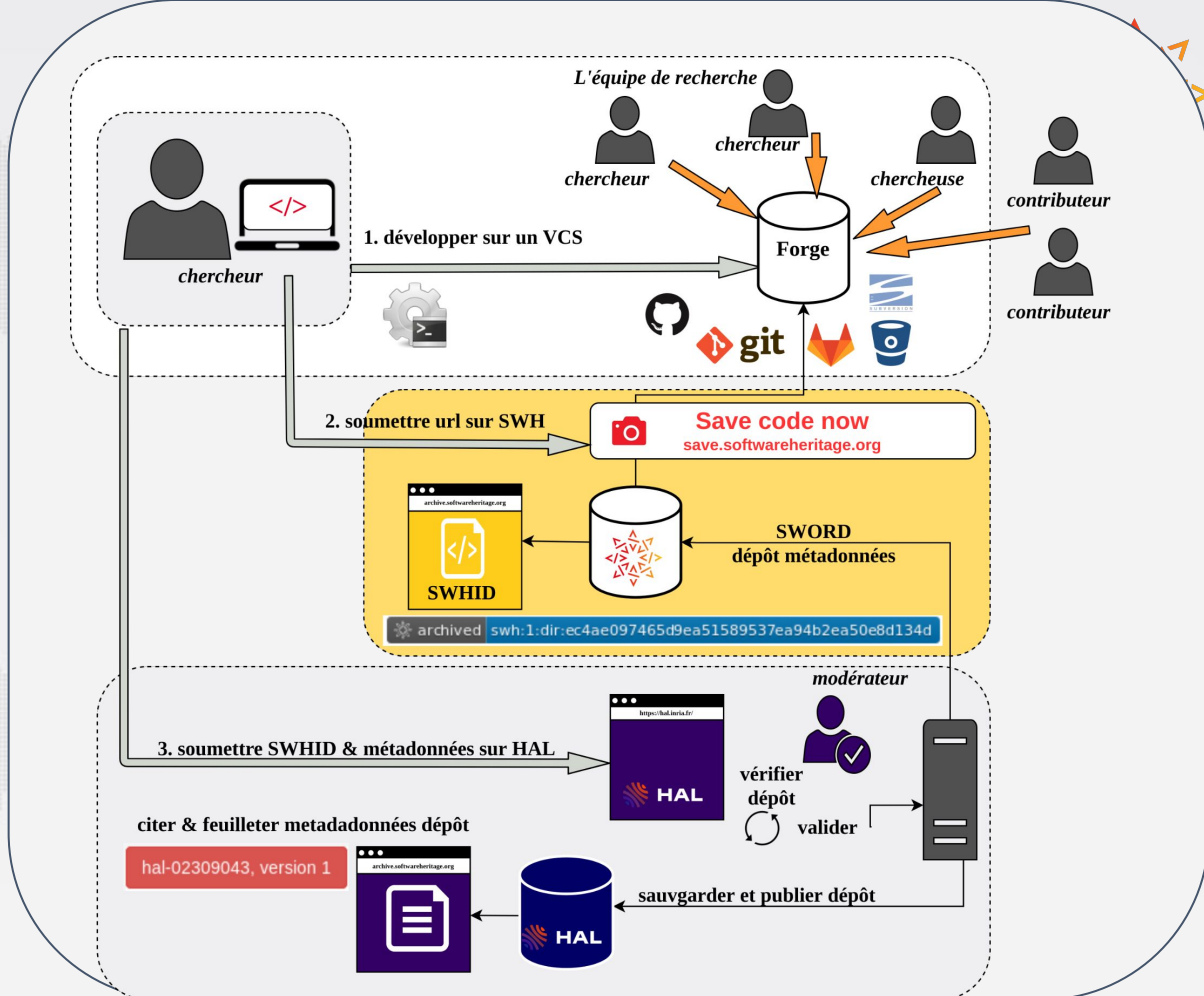
Advantages:

- ★ **Metadata moderation** by the digital archivist team
- ★ Export **formats** available on the software record - to cite software
- ★ Retrieve metadata with **codemeta.json** to complete form



Deposit guide :

Gruenpeter et al. Create software deposit in HAL: User guide and best practices. [Technical Report] Inria; CCSD; Software Heritage. 2022. [\[hal-01872189v2\]](https://hal.archives-ouvertes.fr/hal-01872189v2)





Describe - intrinsic metadata files

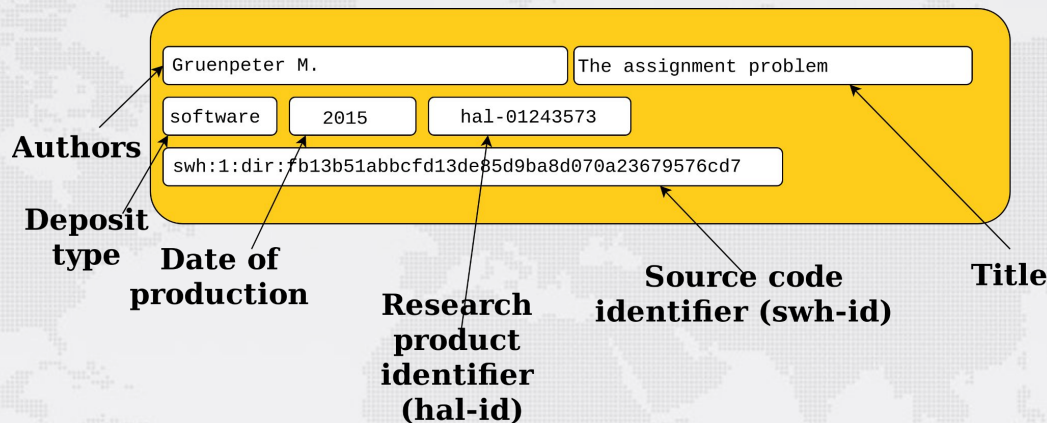
- ★ Prepare your code with the following files :
*These files are verified by **moderators** for a HAL deposit*

- ❑ README (<https://readme.so/fr/editor>)
- ❑ AUTHORS (containing list of authors)
- ❑ LICENSE
 - Open-source [SPDX compliant](https://choosealicense.com/) license
<https://choosealicense.com/>
<https://reuse.software/>
- ❑ **codemeta.json** (not mandatory but useful) - [Generator tool](#)

See also: [HOWTO archive and reference your code](#)



Reference vs. citation



Credit & Attribution

a metadata record
all authors &
contributors

Reuse & Reproducibility

a specific artifact
with complementary
information (docs)



Citation export in HAL

- Citation accessible on the HAL record
- Export BibTeX using the format [BibLaTeX](#) @software or @softwareversion (if a version property was submitted)
- Export used in activity reports for scientific outputs at Inria since 2020.

Softwares

[38] [SW] M. Frigo, M. Zucchelli, R. Deriche and S. Deslauriers-Gauthier, *TALON: Tractograms As Linear Operators in Neuroimaging* version 0.3.0, 19th Jan. 2021. HAL: [\(hal-03116143\)](#), URL: <https://hal.archives-ouvertes.fr/hal-03116143>, VCS: <https://gitlab.inria.fr/cobcom/talon>, SWHID: [\(swh:1:dir:f25157ad1b13cb20ac3457d4f6756b49ac63d079;origin=https://hal.archives-ouvertes.fr/hal-03116143;visit=swh:1:snp:465d89956196578717f4cb5155e456c279aa6a22;anchor=swh:1:rev:10247a14640a280b9140a27ce003d382d70cccac;path=/\)](#).

HAL's citation format

Matteo Frigo, Mauro Zucchelli, Rachid Deriche, Samuel Deslauriers-Gauthier. TALON: Tractograms As Linear Operators in Neuroimaging. 2021, [\(swh:1:dir:f25157ad1b13cb20ac3457d4f6756b49ac63d079;origin=https://hal.archives-ouvertes.fr/hal-03116143;visit=swh:1:snp:465d89956196578717f4cb5155e456c279aa6a22;anchor=swh:1:rev:10247a14640a280b9140a27ce003d382d70cccac;path=/\)](#). [\(hal-03116143\)](#)

```
@softwareversion{frigo:hal-03116143v1,
  TITLE = {{TALON: Tractograms As Linear Operators in
Neuroimaging}},
  AUTHOR = {Frigo, Matteo and Zucchelli, Mauro and
Deriche, Rachid and Deslauriers-Gauthier, Samuel},
  URL = {https://hal.archives-ouvertes.fr/hal-03116143},
  NOTE = {},
  YEAR = {2021},
  MONTH = Jan,
  SWHID =
{swh:1:dir:f25157ad1b13cb20ac3457d4f6756b49ac63d079;origin
=https://hal.archives-ouvertes.fr/hal-03116143;visit=swh:1
:snp:465d89956196578717f4cb5155e456c279aa6a22;anchor=swh:1
:rev:10247a14640a280b9140a27ce003d382d70cccac;path=/},
  VERSION = {0.3.0},
  REPOSITORY = {https://gitlab.inria.fr/cobcom/talon},
  LICENSE = {MIT License},
  KEYWORDS = {diffusion MRI ; dMRI ; tractography ; python
; optimization},
  FILE =
{https://hal.archives-ouvertes.fr/hal-03116143/file/talon-
source.zip},
  HAL_ID = {hal-03116143},
  HAL_VERSION = {v1},
}
```

Just one click- the updateswh browser extension



<https://www.softwareheritage.org/browser-extensions/>

This tab shows the archival status of the repository

Green up to date
Yellow not up to date
Grey: not archived yet
Red: not archivable (private)



Outline

- ★ Preserving source code
- ★ Software Heritage the universal source code archive
- ★ Research Software: a first class research output
- ★ Good practices for software curation
- ★ **Conclusion**



Wrap up

- ★  Archive your code! -Save on SWH and deposit on HAL

<https://save.softwareheritage.org/> [Tutorial](#) [How To](#)

- ★  Describe your code with metadata

[README](#), [LICENSE](#), AUTHORS, [codemeta.json](#)

- ★  Reference your code

SWHID with context

- ★  Cite your code

HAL-ID + SWHID - version, release, file, lines [biblatex-software](#) [Tutorial](#)

The SWH ambassadors program



Who can be an ambassador?

- ★ **Enthusiastic individuals** who wish to volunteer as ambassadors to help grow the [Software Heritage community](#)
- ★ An ambassador can come from **different areas**; academia, cultural heritage, industry and public administration.

Do I need a technical background to become an ambassador?

- ★ Absolutely not! Source code archival is a cross-domain, cross-expertise concern.
- ★ you can have a technical background which might be helpful to read technical documentation and trying some of the more advanced features.



Thank you for your attention!

Keep in touch: morane@softwareheritage.org

@moraneottilia, @SWHeritage

<https://www.softwareheritage.org/newsletter/>

<https://www.softwareheritage.org/community/users/>

<https://www.softwareheritage.org/community/developers/>

<https://www.softwareheritage.org/ambassadors/>



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

