Archive and preserve your source code with Software Heritage

Pierre Poulain
Université de Paris

2021-06-24
iPOP-UP bioinformatic meeting

Software Heritage

Outline

1. Explain why code archiving is important.
2. Archive code in Software Heritage.
3. Apply good practices for code archiving.

Who are you?

WooClap

https://www.wooclap.com/IPOPUP

Embracing reproducibility in bioinformatics

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

EDITORIAL

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, Eivind Hovig

Published: October 24, 2013 • https://doi.org/10.1371/journal.pcbi.1003285

Source: Sandve et al, PLOS Comput Biol, 2013
DOI: 10.1371/journal.pcbi.1003285

Rule 4: Version Control All Custom Scripts

Even the slightest change to a computer program can have large intended or unintended consequences. When a continually developed piece of code (typically a small script) has been used to generate a certain result, only that exact state of the script may be able to produce that exact output, even given the same input data and parameters. As also discussed for rules 3 and 6, exact reproduction of results may in certain situations be essential. If computer code is not systematically archived along its evolution, backtracking to a code state that gave a certain result may be a hopeless task. This can cast doubt on previous results, as it may be impossible to know if they were partly the result of a bug or otherwise unfortunate behavior.

The standard solution to track evolution of code is to use a version control system [15], such as Subversion, Git, or Mercurial. These systems are relatively easy to set up and use, and may be used to systematically store the state of the code throughout development at any desired time granularity.

Using version control systems is good! 👏

Check for updates

OPINION ARTICLE

Four simple recommendations to encourage best practices in research software [version 1; peer review: 3 approved]

Rafael C. Jiménez [1], Mateusz Kuzak[2], Monther Alhamdoosh [3], Michelle Barker[4], Bérénice Batut [5], Mikael Borg[6], Salvador Capella-Gutierrez [7], Neil Chue Hong[8], Martin Cook[1],

Source: Jiménez et al, F1000 Research, 2017
DOI: 10.12688/f1000research.11407.1

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

EDITORIAL

Ten simple rules for making research software more robust

Morgan Taschuk, Greg Wilson

Published: April 13, 2017 • https://doi.org/10.1371/journal.pcbi.1005412

Source: Taschuk & Wilson, PLOS Comput Biol, 2017
DOI: 10.1371/journal.pcbi.1005412

PLOS BIOLOGY

OPEN ACCESS

COMMUNITY PAGE

Best Practices for Scientific Computing

Greg Wilson, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, Kathryn D. Huff, Ian M. Mitchell, Mark D. Plumbley, Ben Waugh, Ethan P. White, Paul Wilson

Published: January 7, 2014 • https://doi.org/10.1371/journal.pbio.1001745

Source: Wilson et al, PLOS Biology, 2014
DOI: 10.1371/journal.pbio.1001745

But wait, who controls your source code on GitHub?

WooClap

https://www.wooclap.com/IPOPUP

GitHub now belongs to Microsoft

Microsoft to acquire GitHub for $7.5 billion

June 4, 2018 | Microsoft News Center

Acquisition will empower developers, accelerate GitHub's growth and advance Microsoft services with new audiences

Source: Microsoft Blog

REUTERS

World Business Markets Breakingviews Video More

TECHNOLOGY, MEDIA & TELECOM - INNOVATION JUNE 5, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

GitLab gains developers after Microsoft buys rival GitHub

By Vibhuti Sharma, Supantha Mukherjee

3 MIN READ

Source: Reuters

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

Google Kills Off Google Code

Natasha Lomas @riptari / 10:58 AM GMT+1 • March 13, 2015

Source: TechCrunch

1.4 million projects

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

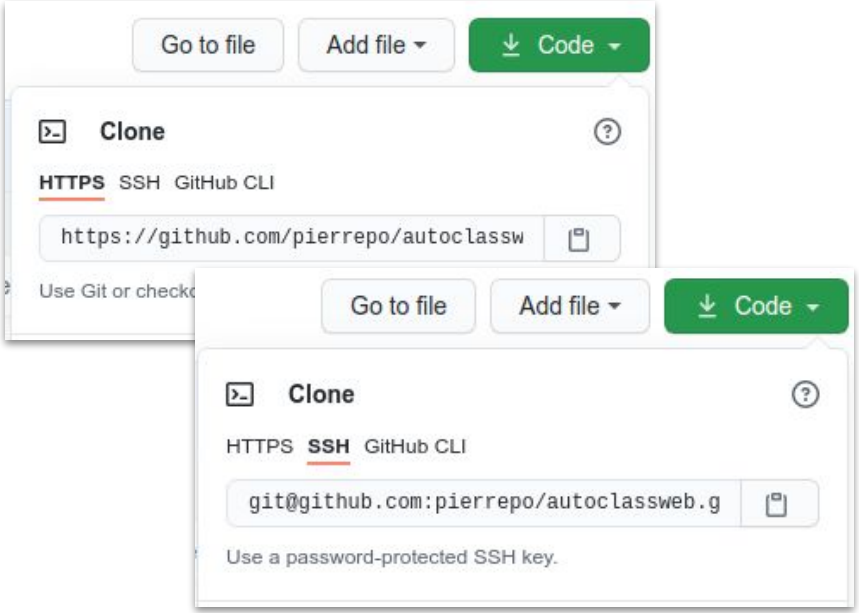
Hosting your open-source project

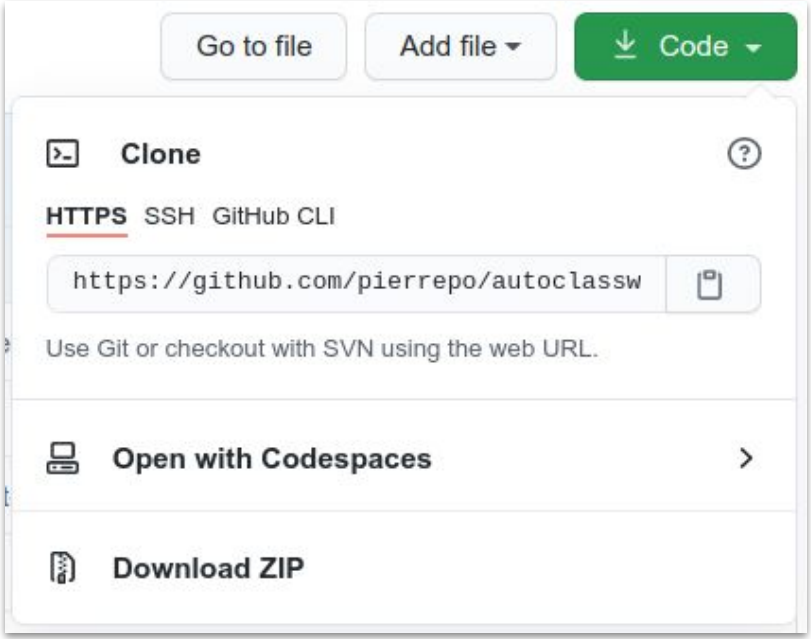
Hosting you (open) source code on a free, publicly available platform is fine.

But you have to prepare for the platform shutdown (you need a plan B).

Source: Geralt, Pixabay

Manual backups






```
git clone ....

git pull 🔄
```

Automated backups

Zenodo

https://zenodo.org/

GitHub integration:
Making Your Code Citable

Archiving to Zenodo is performed automatically but manually triggered by new release on GitHub

Example:
GitHub releases / Zenodo snapshots

Why use Zenodo?

- **Safe** — your research is stored safely for the future in CERN's Data Centre for as long as CERN exists.
- **Trusted** — built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science.
- **Citeable** — every upload is assigned a Digital Object Identifier (DOI), to make them citable and trackable.
- **No waiting time** — Uploads are made available online as soon as you hit publish, and your DOI is registered within seconds.
- **Open or closed** — Share e.g. anonymized clinical trial data with only medical professionals via our restricted access mode.
- **Versioning** — Easily update your dataset with our versioning feature.
- **GitHub integration** — Easily preserve your GitHub repository in Zenodo.
- **Usage statistics** — All uploads display standards compliant usage statistics

Source: Zenodo

Figshare

https://figshare.com/

Belongs to Digital Science

Also provides a DOI

GitHub integration:
How to connect Figshare with your GitHub account

Also triggered by GitHub release

Software Heritage

https://www.softwareheritage.org/

“We are building the universal software archive”

It archives your open-source code permanently and for free.

Disclaimer: I’m Software Heritage ambassador

Software Heritage

Non-profit organization

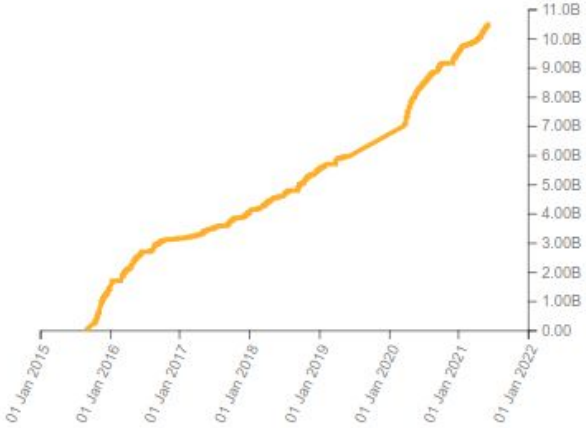
launched in 2016 by INRIA (Roberto Di Cosmo & Stefano Zacchiroli)

supported by UNESCO, the CNRS, Microsoft, Huawei, Intel...

Software Heritage

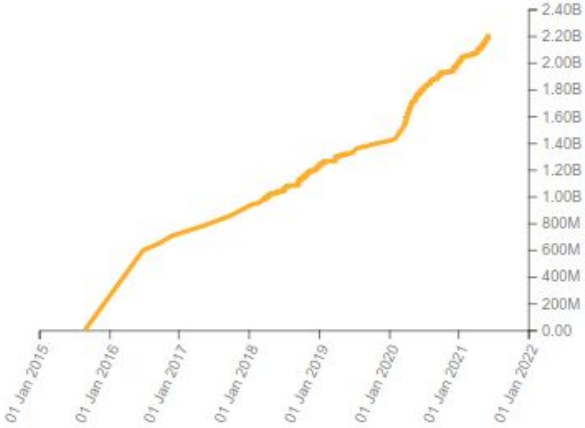
Source files

10 545 239 307



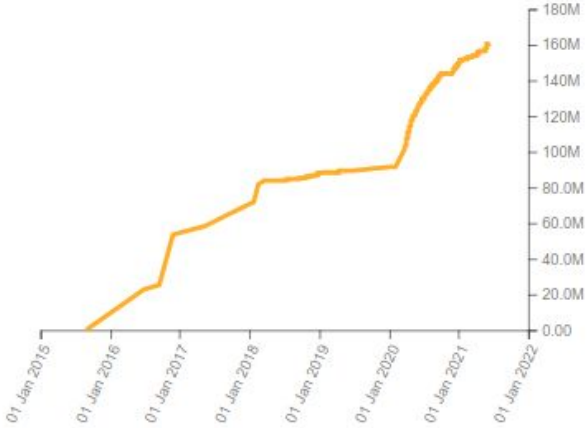
Commits

2 215 028 000



Projects

161 168 065



Directories

8 776 445 689

Authors

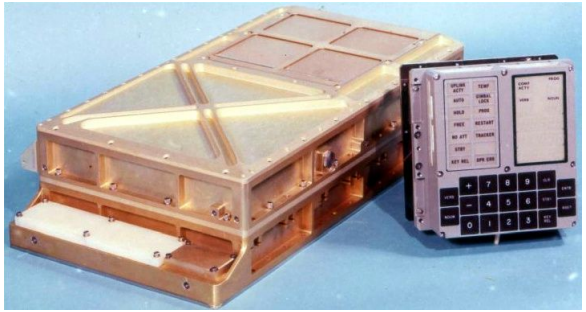
43 822 776

Releases

17 972 096

Source: Software Heritage

Archiving software



Source: Apollo Guidance Computer, NASA, Wikimedia

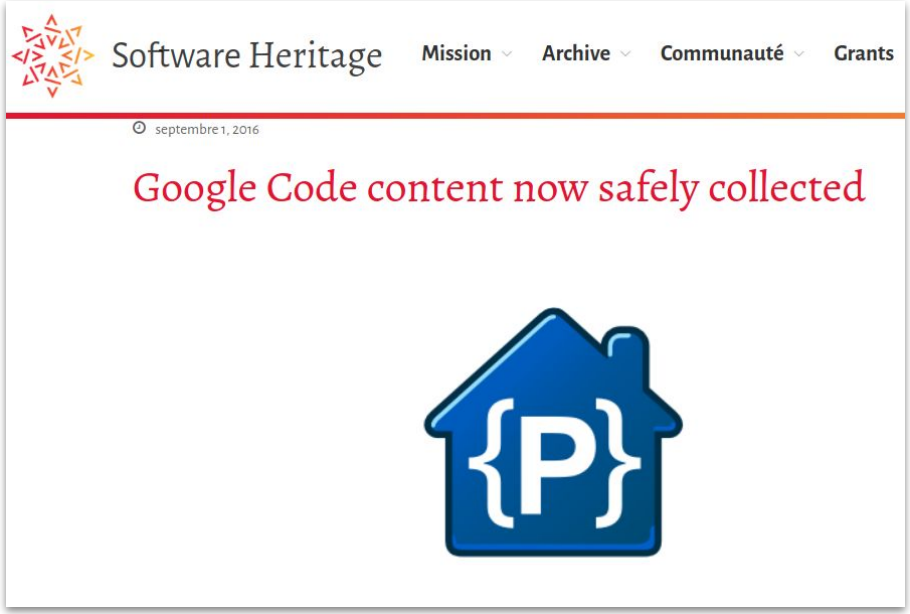
GROMACS
FAST. FLEXIBLE. FREE.

BOW TIE

Archiving software

Source: Software Heritage

Rescuing software

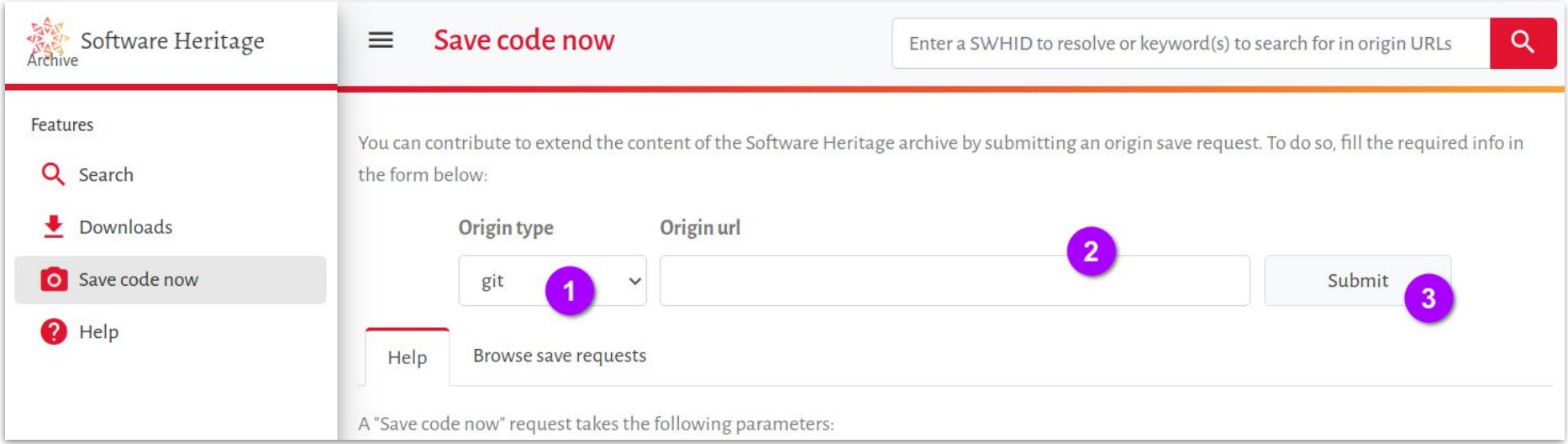


Source: Software Heritage

Source: Software Heritage

Save your code now!

https://archive.softwareheritage.org/save/



But wait, what kind of code? 🤔

WooClap

https://www.wooclap.com/IPOPUP


GOOD NEWS EVERYONE
SOFTWARE HERITAGE
ARCHIVES ANY OPEN-SOURCE CODE!
imgflip.com

Best practices: beyond the code

HOWTO archive and reference your code

Step 1: prepare your public repository

- add a README file
- add an AUTHORS file
- add license information in one of the two recommended ways
  - a LICENSE file at the root of your project, *or*
  - a LICENSES directory containing all the licenses used in your project, and an SPDX compliant copyright header in all your source code files (see the REUSE instructions for details and tools)
- (optionally) add a codemeta.json file containing machine readable metadata (can be produced using the CodeMeta Generator

It is now an accepted practice to also add markdown versions of the README file, but please keep the AUTHORS and LICENSE files as plain text.

Best practices: beyond the code

Metadata for humans

- **`README`**
  https://readme.so/fr/editor

- **`AUTHORS`**
  `Ada Lovelance <ada@programming.org>`
  `Margaret Hamilton <margaret@nasa.com>`

- **`LICENSE`**
  Open-source SPDX compliant license
  https://choosealicense.com/
  https://reuse.software/

Metadata for machines

- **`codemeta.json`**
  with a human-friendly generator

Demo time

Source: **Giphy**

https://archive.softwareheritage.org/save/

https://github.com/patrickfuchs/buildH

Update the archive?

Software Heritage Archive

Browse the archive

Enter a SWHID to resolve or keyword(s) to search for in origin URLs

Features

- Search
- Downloads
- Save code now
- Help

https://github.com/patrickfuchs/buildH

23 June 2021, 15:14 UTC

Code | Branches (1) | Releases (9) | Visits

Branch: HEAD | 660bf0e / | History | Downloa | Save again | Permalinks

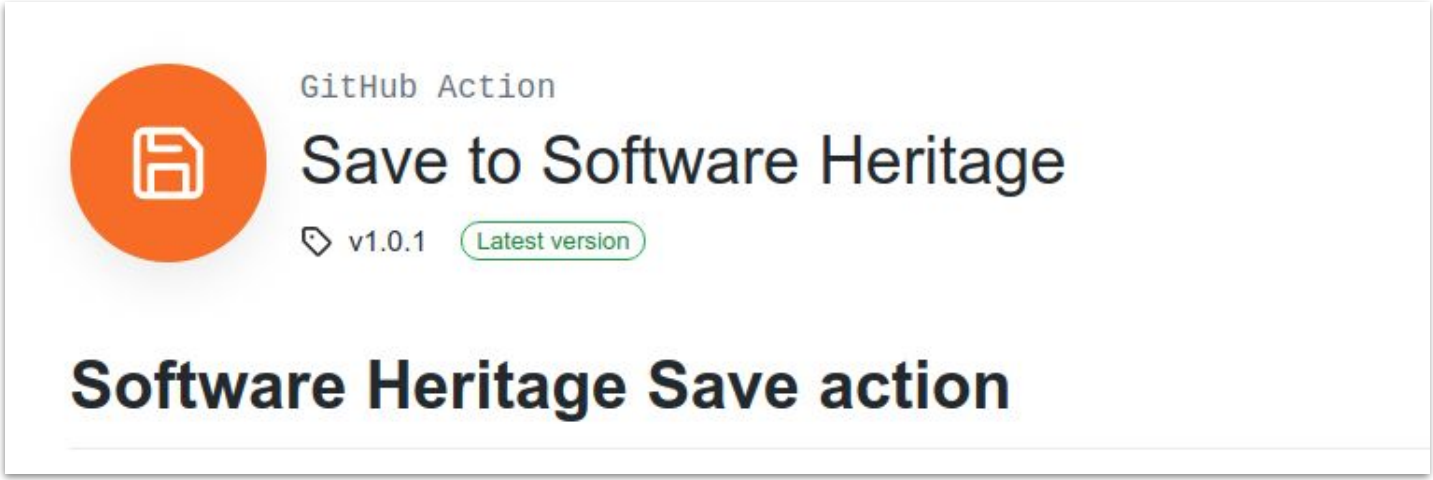
Tip revision: a0b3cd60863c295604e68c718b0d255e25aad8c1 authored by patrickfuchs on 23 June 2021, 13:05 UTC

Remove duplications

File	Mode	Size
.github		
binder		
buildh		

Soon automated by a bot 🤖

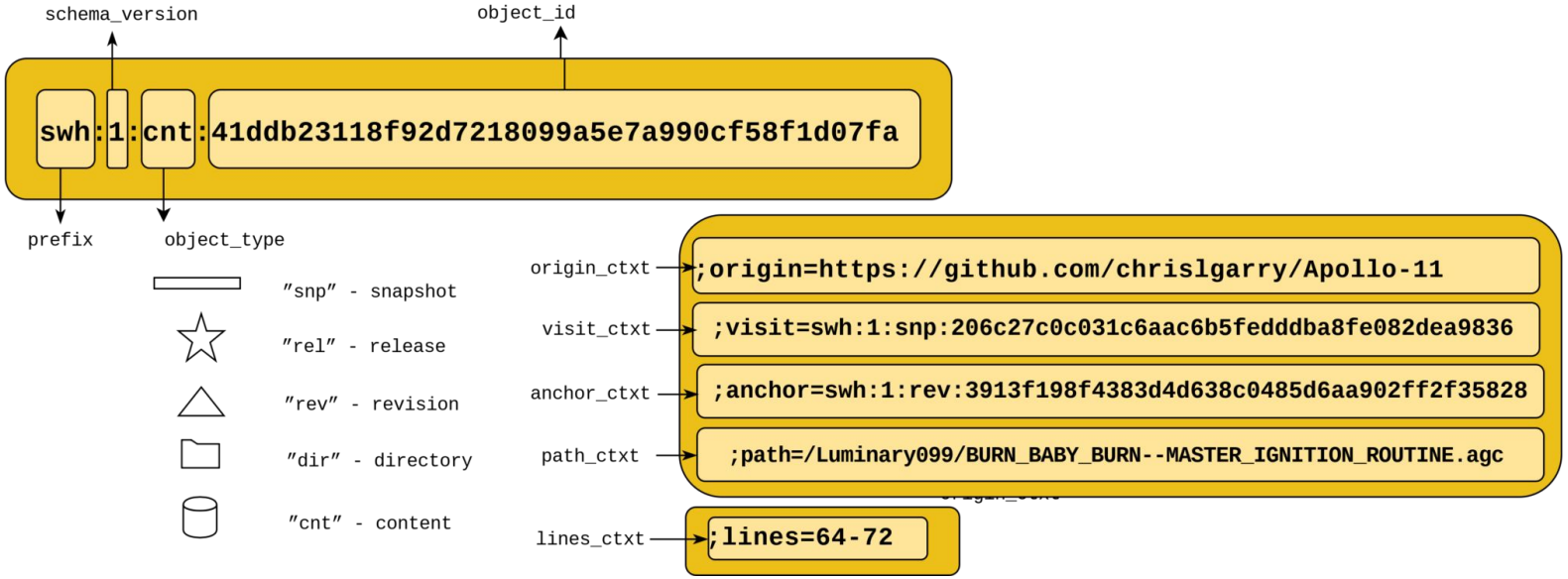
Update the archive?



https://github.com/marketplace/actions/save-to-software-heritage

Reference your software

Use ~~DOI~~ SWHID: an intrinsic persistent identifier



Source: Software Heritage

Reference your software: example

Original GitHub repo: https://github.com/patrickfuchs/buildH

Reference to the archive in Software Heritage (for your README):
https://archive.softwareheritage.org/browse/origin/?origin_url=https://github.com/patrickfuchs/buildH

archived repository

Reference to the archive in Software Heritage (in your paper, for a specific version):

Version 1.3.1: swh:1:dir:c2a73bcfe461c62aee57f9d8ac9af7167e30e7ec

swh:1:dir:c2a73bcfe461c62aee57f9d8ac9af7167e30e7ec;origin=https://github.com/patrickfuchs/buildH;visit=swh:1:snp:7ca072b98e60232bd78ab6a7239778b3b569afdc;anchor=swh:1:rel:97cbf640f826e1272bad00e7e3438c48566db184

Cite your software archive

BibLaTex style extension for software

[Software Release] B. Langmead and S. L. Salzberg, *Bowtie2* version 2.4.2, Oct. 2022. LIC: GPL. URL: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml, VCS: https://github.com/BenLangmead/bowtie2, SWHID: ⟨swh:1:rel:97bacffeaa6e7c3f574ce5b566daba82aa18a11f;origin=https://github.com/BenLangmead/bowtie2;visit=swh:1:snp:c25778cfefc086c63c6f78eed230d0b9c88876ee⟩.

[Software excerpt] MIT Instrumentation Laboratory, "AGC Luminary routine for changing LEM asset during landing", from *Apollo 11 Guidance Computer (AGC) source code for the command and lunar module* 1967. VirtualAGC project. LIC: Public Domain. URL: https://www.ibiblio.org/apollo, VCS: https://github.com/virtualagc/virtualagc, SWHID: ⟨swh:1:cnt:64582b78792cd6c2d67d35da5a11bb80886a6409;origin=https://github.com/virtualagc/virtualagc;anchor=swh:1:rev:007c2b95f301f9438b8b74d7993b7a3b9a66255b;lines=245-261⟩.

Wrap-up

📦 Archive your code!

https://archive.softwareheritage.org/save/

HAL archives-ouvertes.fr

👓 Describe your code with metadata

README, LICENSE, AUTHORS, codemeta.json

🗺️ Reference your code

SWHID over DOI, context

👩‍🎓 Cite your code

Version, release, file, lines