Software Heritage, l'archive mondiale du code source logiciel

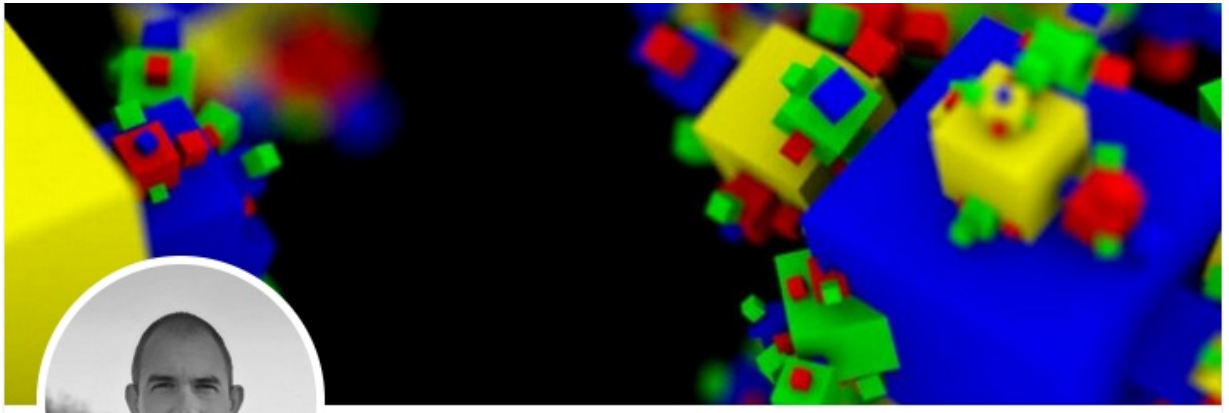
Pierre Poulain

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Les Rendez-vous du Centre des Humanités Numériques | 2022-04-21

Bonjour 👋



Pierre Poulain
@pierrepo

Bioinformatics. Python. Unix. Associate prof at @Univ_Paris and @IJMonod #proteomics #openscience #opensource @SWHeritage ambassador #pedagogy #Certifiens

Paris, France cupnet.net

pierre.poulain@u-paris.fr

@pierrepo

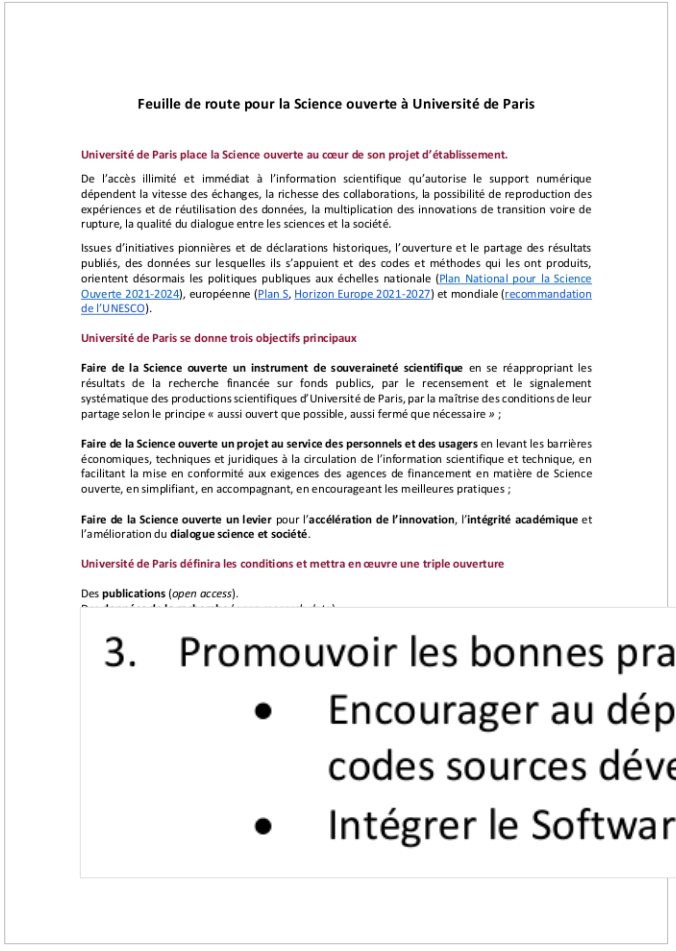
Software Heritage ?

Source : Ouvrir la science, 2021

→ Soutenir **Software Heritage** et recommander son adoption pour l'archivage et le référencement des codes sources.

→ Proposer la standardisation du **Software Heritage Identifier (SWHID),** qui complètera les DOI pour les logiciels.

Software Heritage ?



Feuille de route pour la Science ouverte à Université de Paris

Université de Paris place la Science ouverte au cœur de son projet d'établissement.

De l'accès illimité et immédiat à l'information scientifique qu'autorise le support numérique dépendent la vitesse des échanges, la richesse des collaborations, la possibilité de reproduction des expériences et de réutilisation des données, la multiplication des innovations de transition voire de rupture, la qualité du dialogue entre les sciences et la société.

Issues d'initiatives pionnières et de déclarations historiques, l'ouverture et le partage des résultats publiés, des données sur lesquelles ils s'appuient et des codes et méthodes qui les ont produits, orientent désormais les politiques publiques aux échelles nationale (Plan National pour la Science Ouverte 2021-2024), européenne (Plan S, Horizon Europe 2021-2027) et mondiale (recommandation de l'UNESCO).

Université de Paris se donne trois objectifs principaux

Faire de la Science ouverte un instrument de souveraineté scientifique en se réappropriant les résultats de la recherche financée sur fonds publics, par le recensement et le signalement systématique des productions scientifiques d'Université de Paris, par la maîtrise des conditions de leur partage selon le principe « aussi ouvert que possible, aussi fermé que nécessaire » ;

Faire de la Science ouverte un projet au service des personnels et des usagers en levant les barrières économiques, techniques et juridiques à la circulation de l'information scientifique et technique, en facilitant la mise en conformité aux exigences des agences de financement en matière de Science ouverte, en simplifiant, en accompagnant, en encourageant les meilleures pratiques ;

Faire de la Science ouverte un levier pour l'**accélération de l'innovation**, l'**intégrité académique** et l'amélioration du **dialogue science et société**.

Université de Paris définira les conditions et mettra en œuvre une triple ouverture

Des **publications** (*open access*).

3. Promouvoir les bonnes pratiques en lien avec l'utilisation de la plateforme Software Heritage
    - Encourager au dépôt dans Software Heritage, en particulier via HAL, des logiciels et des codes sources développés au sein d'Université de Paris
    - Intégrer le Software Heritage Identifier (SWHID) à la politique sur les identifiants ouverts.

Source : Université Paris Cité, 2021

Au menu

1. Pourquoi l'archivage des codes sources des logiciels est important ?

2. Qu'est-ce que Software Heritage ?

3. Comment archiver son code source dans Software Heritage ?

4. Quelles sont les bonnes pratiques pour archiver un code source ?

Cette présentation est garantie sans ligne de code

Les logiciels sont partout !

NEWS FEATURE

THE TOP 100 PAPERS

Nature *explores the most-cited research of all time.*

BY RICHARD VAN NOORDEN, BRENDAN MAHER AND REGINA NUZZO

The discovery of high-temperature superconductors, the determination of DNA's double-helix structure, the first observations that the expansion of the Universe is accelerating — all of these breakthroughs won Nobel prizes and international acclaim. Yet none of the papers that announced them comes anywhere close to ranking among the 100 most highly cited papers of all time.

Citations, in which one paper refers to earlier works, are the standard means by which authors acknowledge the source of their methods, ideas and findings, and are often used as a rough measure of a paper's importance. Fifty years ago, Eugene Garfield published the Science Citation Index (SCI), the first systematic effort to track citations in the scientific literature. To mark the anniversary, *Nature* asked Thomson Reuters, which now owns the SCI, to list the 100 most highly cited papers of all time. (See the full list at www.nature.com/top100.) The search covered all of Thomson Reuter's Web of Science, an online version of the SCI that also includes databases covering the social sciences, arts and humanities, conference proceedings and some books. It lists papers published from 1900 to the present day.

The exercise revealed some surprises, not least that it takes a staggering 12,119 citations to rank in the top 100 — and that many of the world's most famous papers do not make the cut. A few that do, such as the first observation[1] of carbon nanotubes (number 36) are indeed classic discoveries. But the vast majority describe experimental methods or software that have become essential in their fields.

The most cited work in history, for example, is a 1951 paper[2] describing an assay to determine the amount of protein in a solution. It has now gathered more than 305,000 citations — a recognition that always puzzled its lead author, the late US biochemist Oliver Lowry. "Although I really know it is not a great paper … I secretly get a kick out of the response," he wrote in 1977.

The colossal size of the scholarly literature means that the top-100 papers are extreme outliers. Thomson Reuter's Web of Science holds some 58 million items. If that corpus were scaled to Mount Kilimanjaro, then the 100 most-cited papers would represent just 1 centimetre at the peak. Only 14,499 papers — roughly a metre and a half's worth — have more than 1,000 citations (see 'The paper mountain'). Meanwhile, the foothills comprise works that have been cited only once, if at all — a group that encompasses roughly half of the items.

Nobody fully understands what distinguishes the sliver at the top from papers that are merely very well known — but researchers' customs explain some of it. Paul Wouters, director of the Centre for Science and Technology Studies in Leiden, the Netherlands, says that many methods papers "become a standard reference that one cites in order to make clear to other scientists what kind of work one is doing". Another common practice in science ensures that truly foundational discoveries — Einstein's special theory of relativity, for instance — get fewer citations than they might deserve: they are so important that they quickly enter the textbooks or are incorporated into the main text of papers as terms deemed so familiar that they do not need a citation.

Citation counts are riddled with other confounding factors. The volume of citations has increased, for example — yet older papers have had more time to accrue citations. Biologists tend to cite one another's work more frequently than, say, physicists. And not all fields produce the same number of publications. Modern bibliometricians therefore recoil from methods as crude as simply counting citations when they want to measure a paper's value: instead, they prefer to compare counts for papers of similar age, and in comparable fields.

Nor is Thomson Reuters' list the only ranking system available. Google Scholar compiled its own top-100 list for *Nature*. It is based on many more citations because the search engine culls references from a much greater (although poorly characterized) literature base, including from a large range of books. In that list, available at www.nature.com/top100, economics papers have more prominence. Google Scholar's list also features books, which Thomson Reuters did not analyse. But among the science papers, many of the same titles show up.

Yet even with all the caveats, the old-fashioned hall of fame still has value. If nothing else, it serves as a reminder of the nature of scientific knowledge. To make exciting advances, researchers rely on relatively unsung papers to describe experimental methods, databases and software.

Here *Nature* tours some of the key methods that tens of thousands of citations have hoisted to the top of science's Kilimanjaro — essential, but rarely thrust into the limelight.

BIOLOGICAL TECHNIQUES

For decades, the top-100 list has been dominated by protein biochemistry. The 1951 paper[2] describing the Lowry method for quantifying protein remains practically unreachable at number 1, even though many biochemists say that it and the competing Bradford assay[3] — described by paper number 3 on the list — are a tad outdated. In between, at number 2, is Laemmli buffer[4], which is used in a different kind of protein analysis. The dominance of these techniques is attributable to the high volume of citations in cell and molecular biology, where they remain indispensable tools.

At least two of the biological techniques described by top-100 papers have resulted in Nobel prizes. Number 4 on the list describes the DNA-sequencing method[5] that earned the late Frederick Sanger his share of the 1980 Nobel Prize in Chemistry. Number 63 describes polymerase chain reaction ▶

PHOTO BY KYLE BEAN; DESIGN BY WESLEY FERNANDES/NATURE

550 | NATURE | VOL 514 | 30 OCTOBER 2014

« *But the vast majority describe experimental methods or* ***software that have become essential in their fields****.* »

Source: van Noorden et al, Nature, 2014
DOI: 10.1038/514550a

Science ouverte: montre-moi ton code !

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

ACS EDITORS' CHOICE

pubs.acs.org/jcim Viewpoint

Code Sharing in the Open Science Era

W. Patrick Walters*

Cite This: *J. Chem. Inf. Model.* 2020, 60, 4417–4420

Read Online

ACCESS | Metrics & More | Article Recommendations

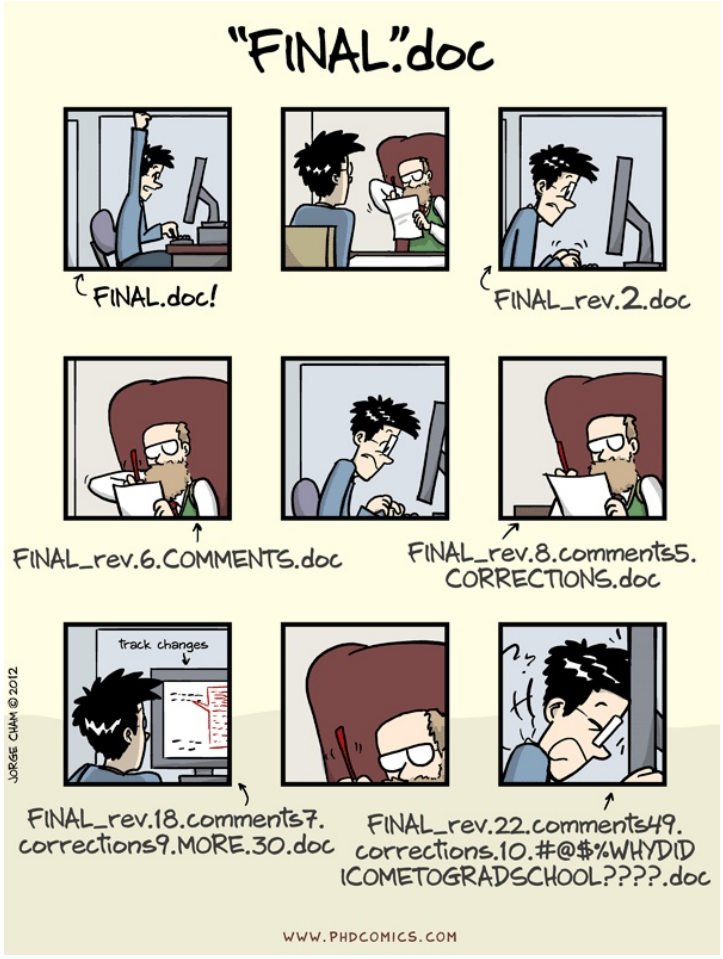
ABSTRACT: Many high-profile scientific journals have established policies mandating the release of code accompanying papers that describe computational methods. Unfortunately, the majority of journals that publish papers in Computational Chemistry and Cheminformatics have yet to define such guidelines. This Viewpoint reviews the current state of reproducibility for the field and makes a case for the inclusion of code with computational papers.

Source: Walters, JCIM, 2020, DOI: 10.1021/acs.jcim.0c01000

Les 3 piliers de la science ouverte

- Libre accès aux publications scientiques (*open access*)
- Données ouvertes (*open data*)
- Codes logiciels ouverts (*open source*)

L'essor des gestionnaires de versions



Source : PhD Comics

OPINION ARTICLE

Check for updates

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 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

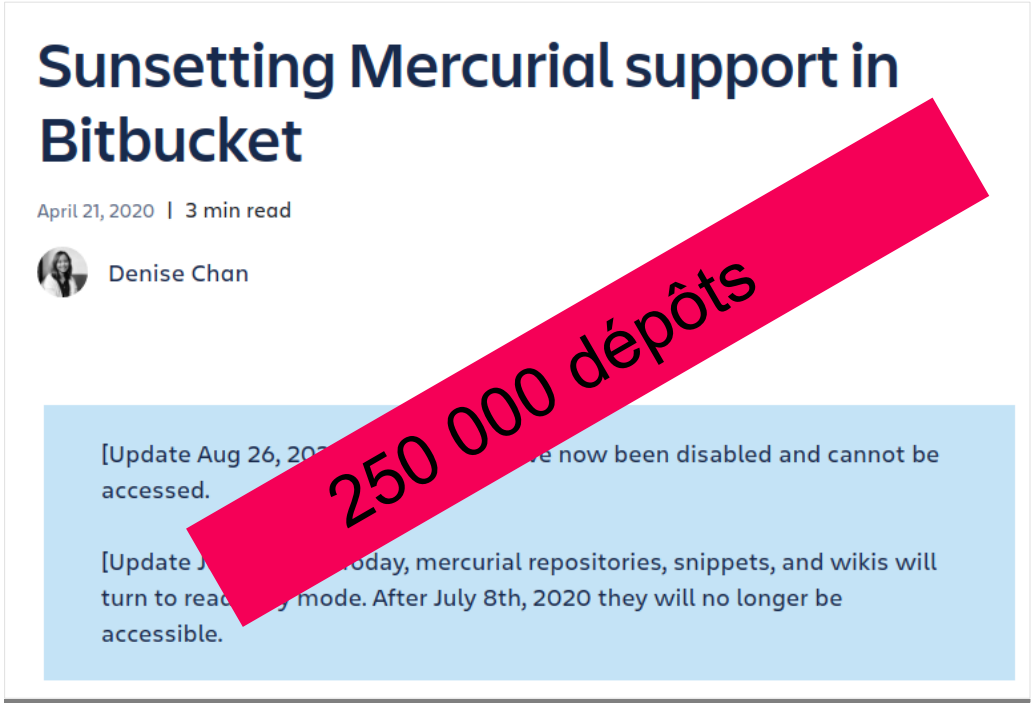
Published: January 7, 2014 • https://doi.org/10.1371/journal.pbio.1001745

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

et des plateformes de développement gratuites

et des plateformes de développement ~~gratuites~~ disparues

Source: Tech



Source: BitBucket blog

La reproductibilité en science nécessite un accès au code source sur le long terme

Héberger votre code sur une plateforme publique et gratuite est acceptable.

Mais vous devez vous préparer à la fermeture de cette plateforme !

Source: Geralt, Pixabay

Archiver du code comme des données ?

Zenodo

OpenAIRE + CERN

Figshare

Digital Science

Code citable avec un DOI

Archivage automatisable depuis GitHub

Archivez votre code sur Software Heritage

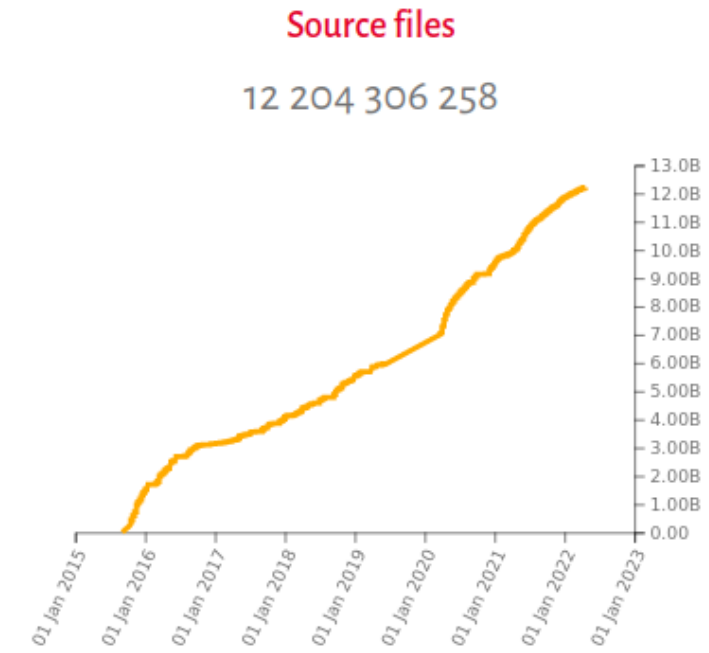
THE GREAT LIBRARY OF SOURCE CODE

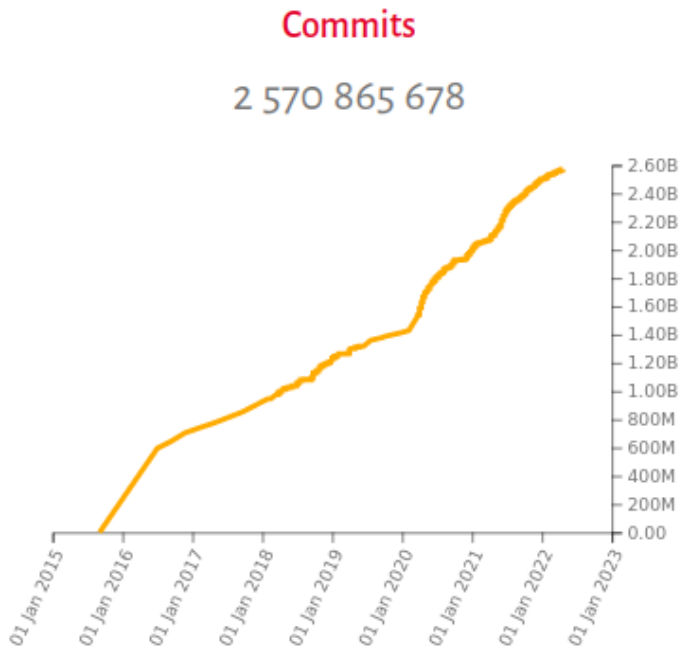
Software Heritage archive tous les codes sources ouverts, pour toujours et gratuitement

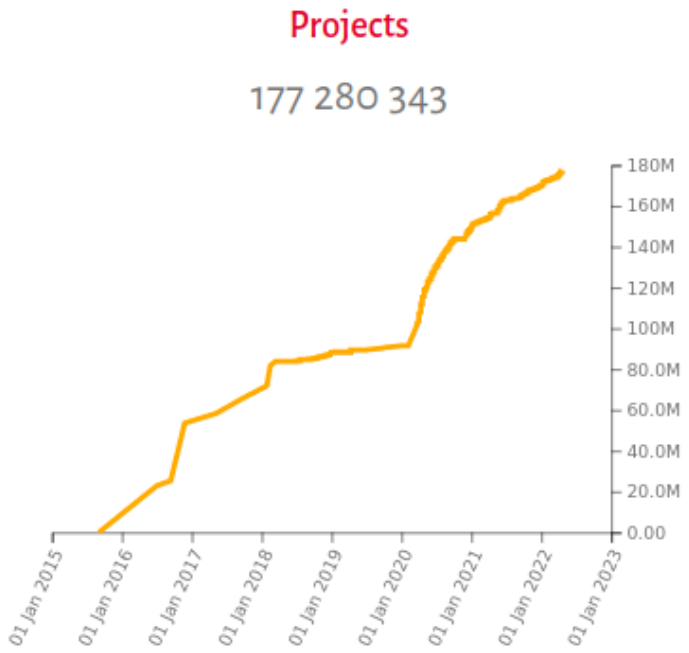
- Organisation à but non lucratif
- Créée en 2016 à l'INRIA (Roberto Di Cosmo & Stefano Zacchiroli)
- Financée par l'UNESCO, le CNRS, UPC, Microsoft, Google, Huawei, Intel...

Source: Software Heritage 5 years anniversary (2021)

Une archive conséquente







Source: Software Heritage

qui intègre déjà de nombreuses « sources »

Bitbucket
2,074,834 origins

git
22,235 origins

R
19,841 origins

debian
126,944 origins

5,875 origins

GitHub
132,824,442 origins

GitLab
4,082,399 origins

Guix
11,821 origins

GNU
354 origins

heptapod
1,039 origins

launchpad
20,417 origins

NixOS
11,821 origins

npm
1,802,916 origins

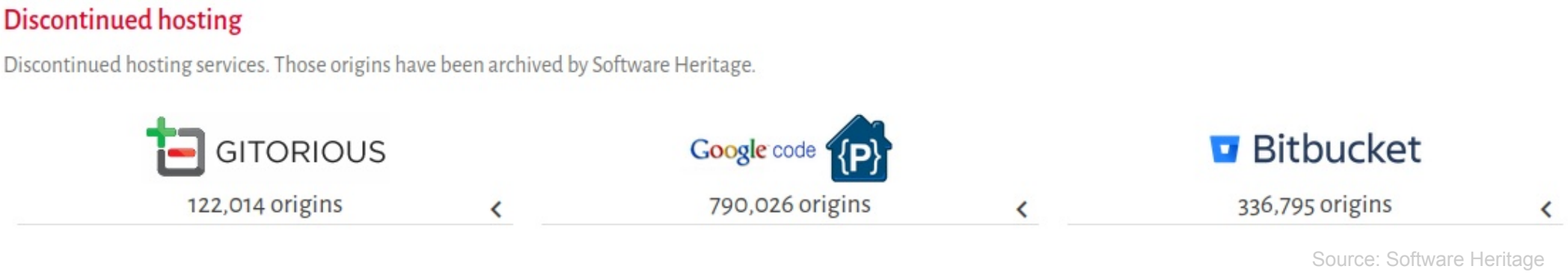
4,083 origins

python Package Index
462,245 origins

SOURCEFORGE
313,582 origins

Source: Software Heritage

Software Heritage sauve les codes sources

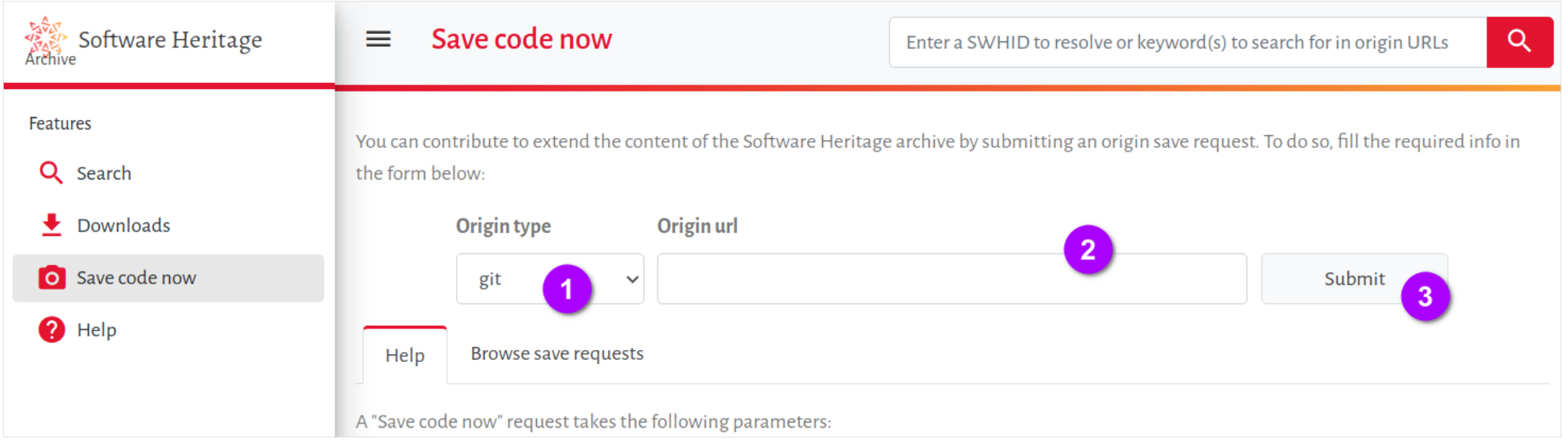


« Google Code content now safely collected », 2016

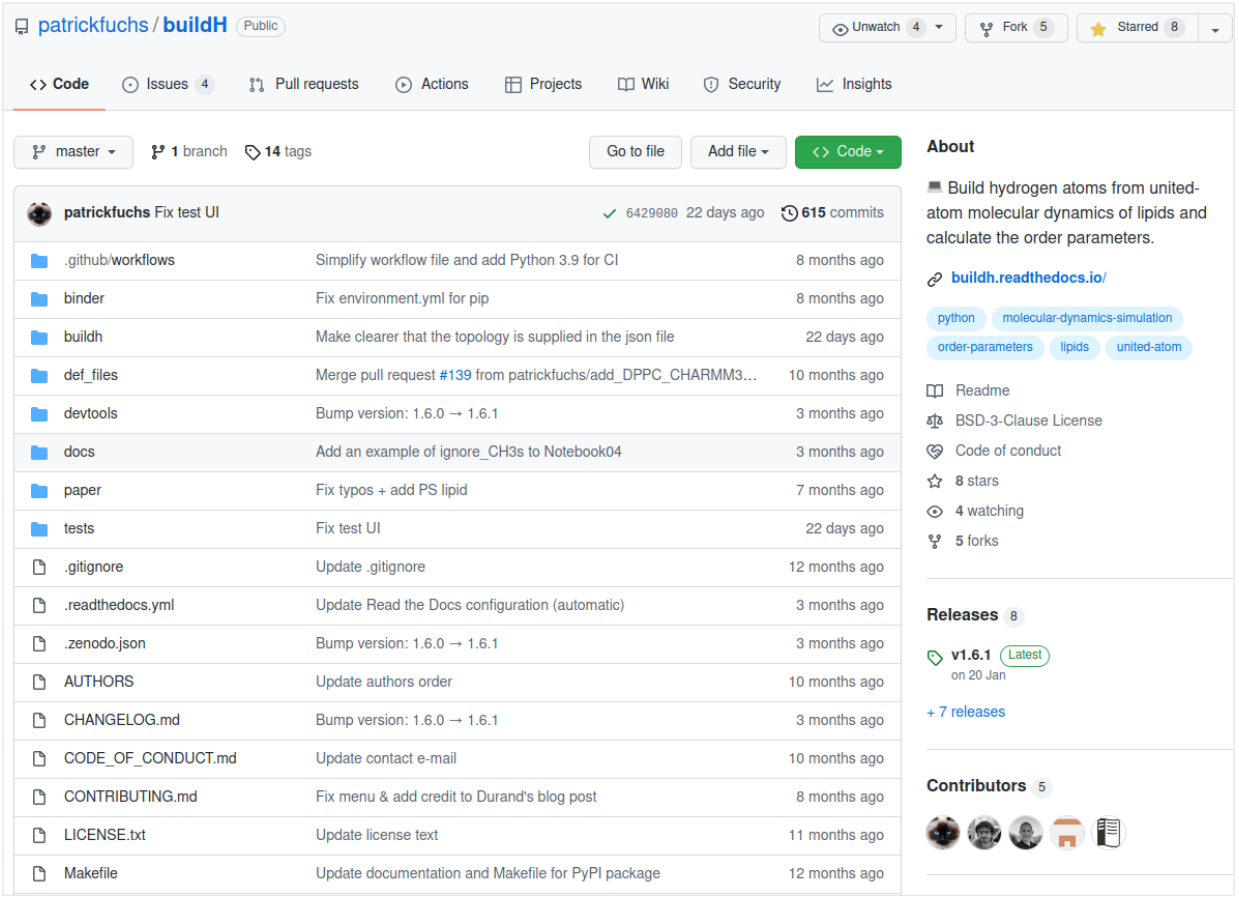
« Rescuing 250000+ endangered Mercurial repositories », 2020

Sauvegardez votre code maintenant !

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

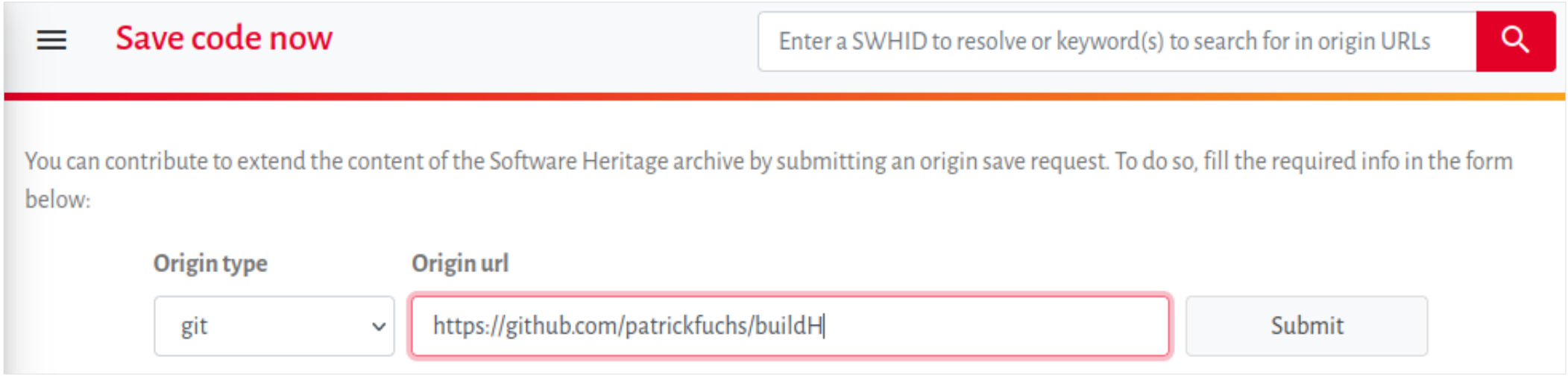


Sauvegarder mon code

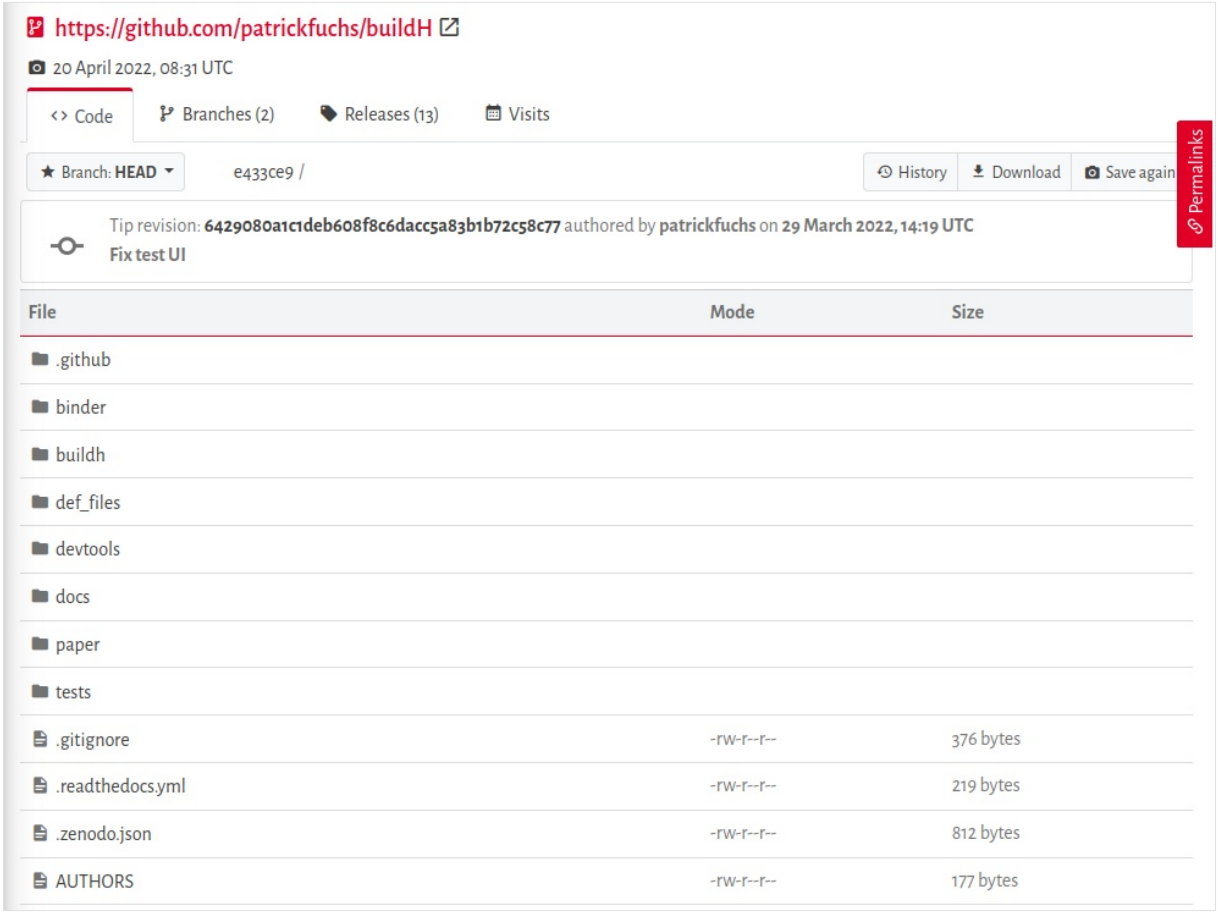


https://github.com/patrickfuchs/buildH

Sauvegarder mon code



Sauvegarder mon code



https://archive.softwareheritage.org/browse/origin/directory/?origin_url=https://github.com/patrickfuchs/buildH


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

Bonnes pratiques : au-delà du code

Metadonnées pour les humains

- **`README`**
  https://readme.so/fr/editor
- **`AUTHORS`**
  ```
  Ada Lovelance <ada@programming.org>
  Margaret Hamilton <margaret@nasa.com>
  ```
- **`LICENSE`**
  Licence ouverte compatible SPDX
  https://choosealicense.com/
  https://reuse.software/

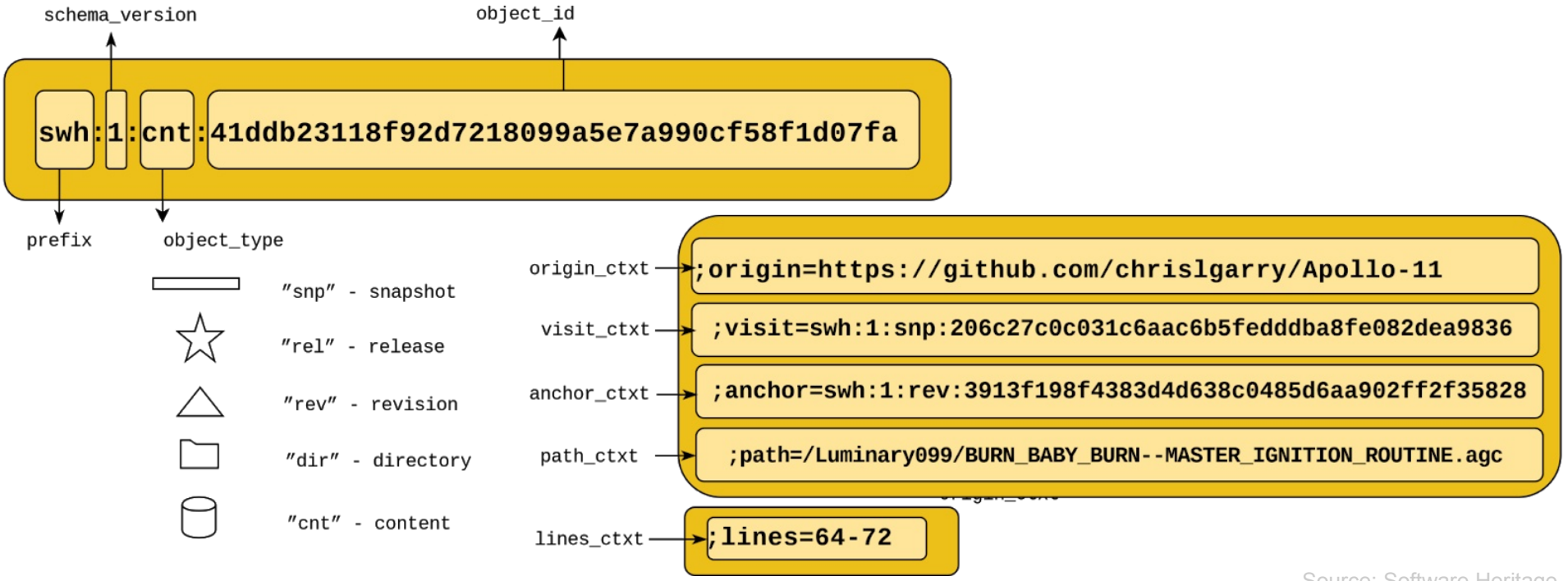
Metadonnées pour les machines

- **`codemeta.json`**
  avec un générateur pour
  les humains : CodeMeta Generator

Voir aussi « HOWTO archive and reference your code »

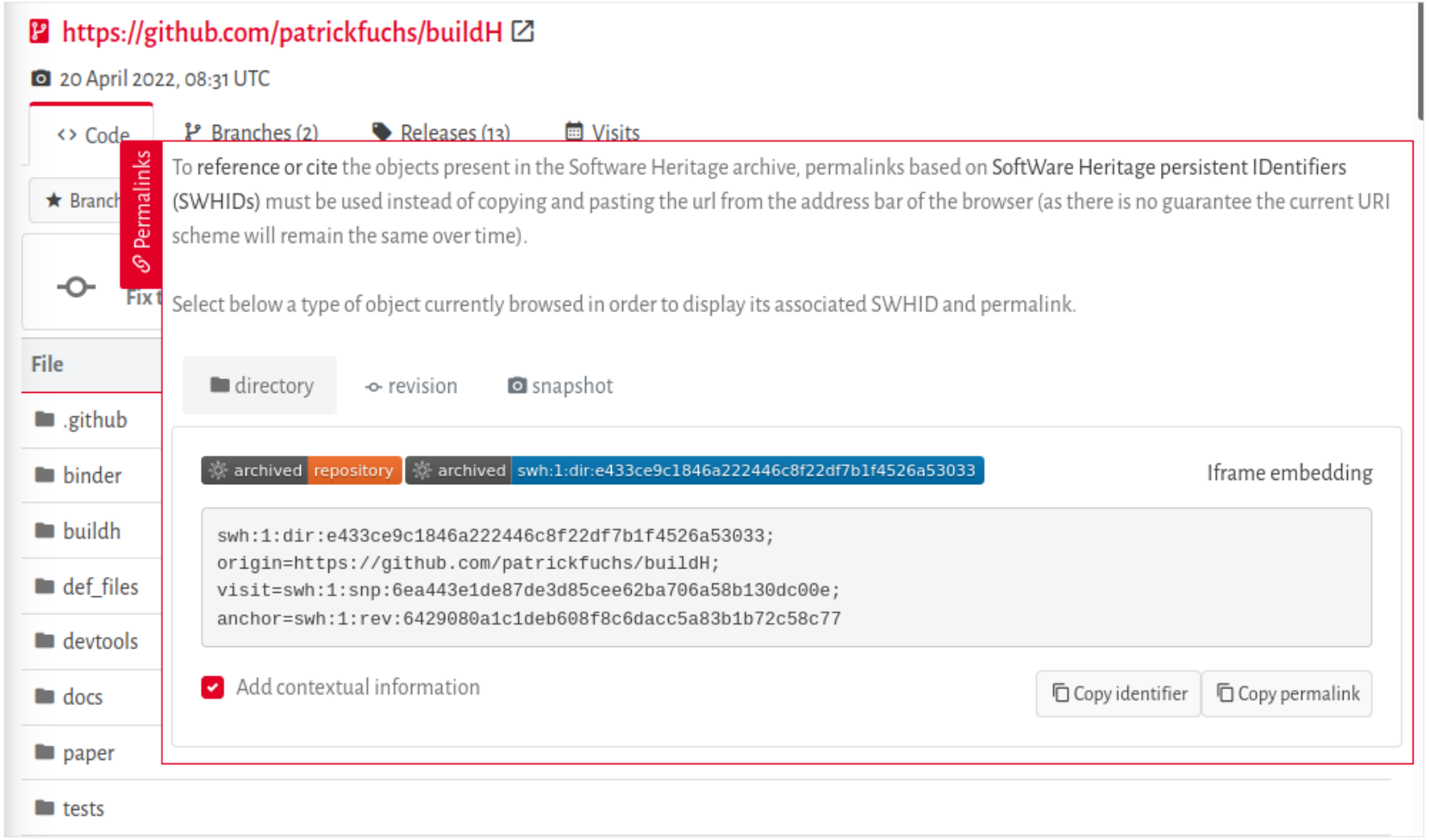
Référencer son code source

Le ~~DOI~~ SWHID : un identifiant intrinsèque et persistant



Source: Software Heritage

Référencer son code source



Référencer son code source

L'adresse du dépôt sur GitHub :
https://github.com/patrickfuchs/buildH

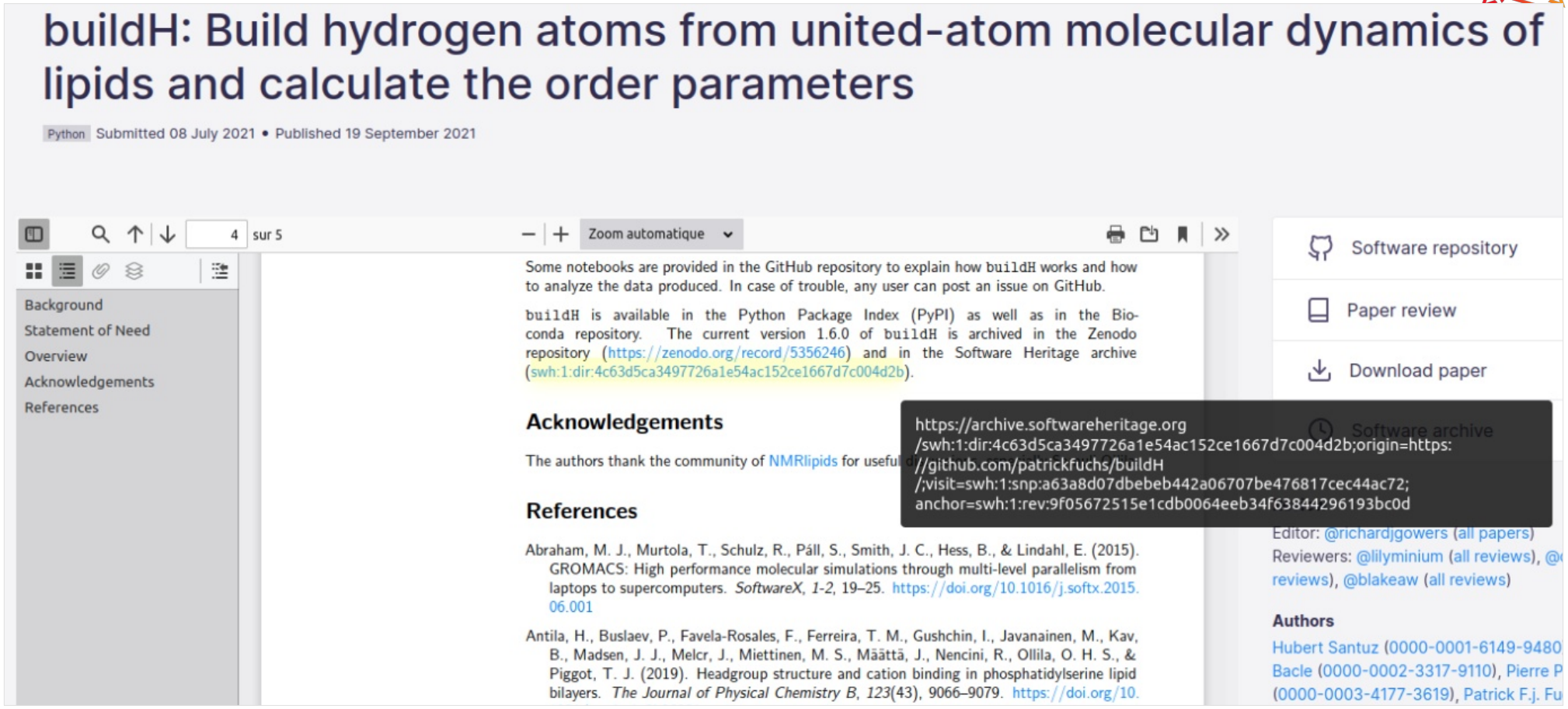
La référence de l'archive dans SWH pour le README : 

https://archive.softwareheritage.org/browse/origin/directory/?origin_url=https://github.com/patrickfuchs/buildH

La référence de l'archive dans SWH pour un article scientifique, avec une version spécique :

```
swh:1:dir:e433ce9c1846a222446c8f22df7b1f4526a53033;
origin=https://github.com/patrickfuchs/buildH;
visit=swh:1:snp:6ea443e1de87de3d85cee62ba706a58b130dc00e;
anchor=swh:1:rev:6429080a1c1deb608f8c6dacc5a83b1b72c58c77
```

Référencer son code source



buildH: Build hydrogen atoms from united-atom molecular dynamics of lipids and calculate the order parameters

Python Submitted 08 July 2021 • Published 19 September 2021

4 sur 5 — Zoom automatique

Background
Statement of Need
Overview
Acknowledgements
References

Some notebooks are provided in the GitHub repository to explain how buildH works and how to analyze the data produced. In case of trouble, any user can post an issue on GitHub.

buildH is available in the Python Package Index (PyPI) as well as in the Bio-conda repository. The current version 1.6.0 of buildH is archived in the Zenodo repository (https://zenodo.org/record/5356246) and in the Software Heritage archive (swh:1:dir:4c63d5ca3497726a1e54ac152ce1667d7c004d2b).

Acknowledgements

The authors thank the community of NMRlipids for useful

https://archive.softwareheritage.org/swh:1:dir:4c63d5ca3497726a1e54ac152ce1667d7c004d2b;origin=https://github.com/patrickfuchs/buildH/;visit=swh:1:snp:a63a8d07dbebeb442a06707be476817cec44ac72;anchor=swh:1:rev:9f05672515e1cdb0064eeb34f63844296193bc0d

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1-2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

Antila, H., Buslaev, P., Favela-Rosales, F., Ferreira, T. M., Gushchin, I., Javanainen, M., Kav, B., Madsen, J. J., Melcr, J., Miettinen, M. S., Määttä, J., Nencini, R., Ollila, O. H. S., & Piggot, T. J. (2019). Headgroup structure and cation binding in phosphatidylserine lipid bilayers. *The Journal of Physical Chemistry B*, *123*(43), 9066–9079. https://doi.org/10.

Software repository
Paper review
Download paper
Software archive

Editor: @richardjgowers (all papers)
Reviewers: @lilyminium (all reviews), @ reviews), @blakeaw (all reviews)

Authors

Hubert Santuz (0000-0001-6149-9480
Bacle (0000-0002-3317-9110), Pierre P
(0000-0003-4177-3619), Patrick F.J. Fu

https://joss.theoj.org/papers/10.21105/joss.03521

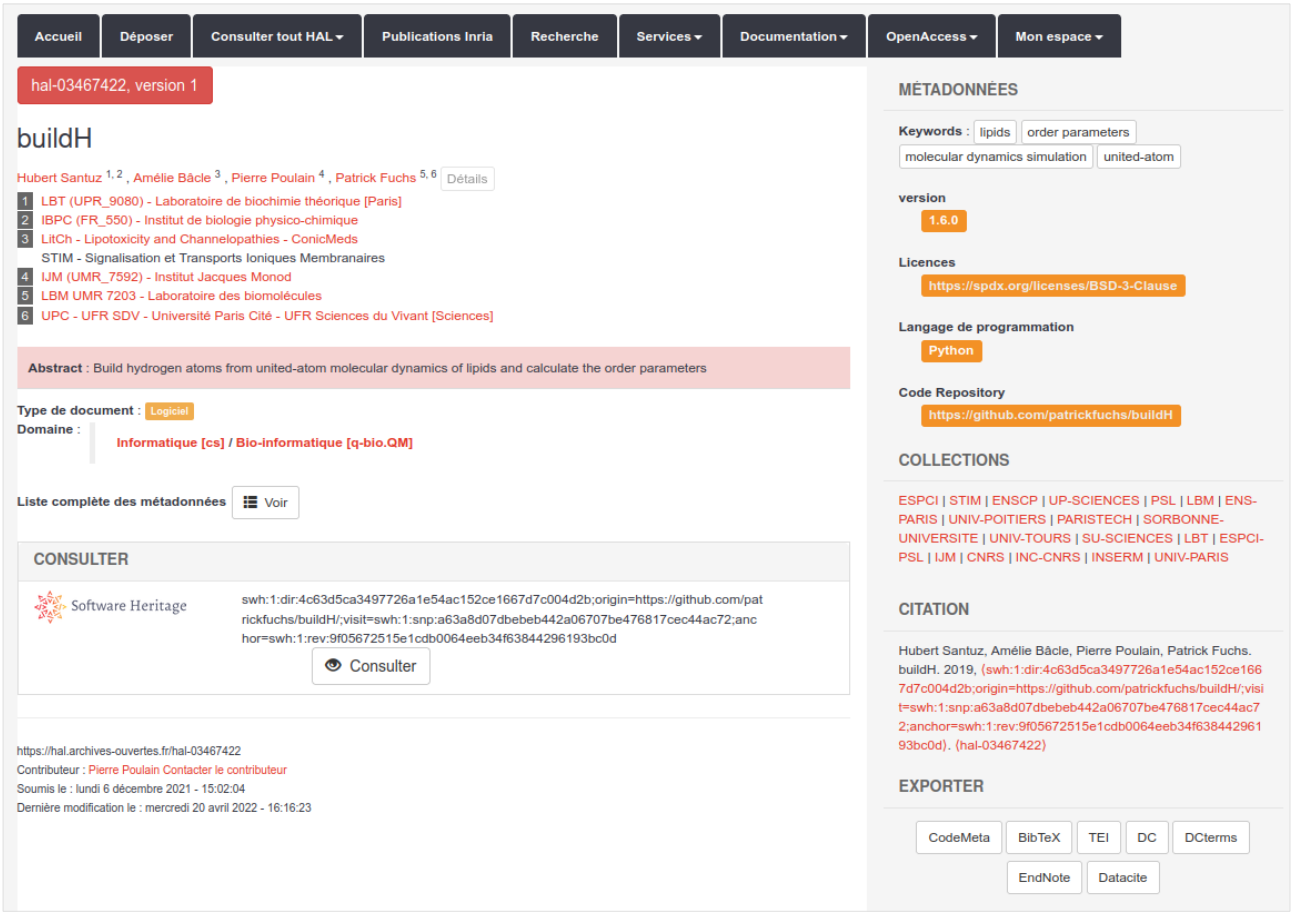
Citer son code source

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⟩.

Citer son code source



Accueil | Déposer | Consulter tout HAL | Publications Inria | Recherche | Services | Documentation | OpenAccess | Mon espace

hal-03467422, version 1

buildH

Hubert Santuz [1, 2], Amélie Bâcle [3], Pierre Poulain [4], Patrick Fuchs [5, 6] Détails

1 LBT (UPR_9080) - Laboratoire de biochimie théorique [Paris]
2 IBPC (FR_550) - Institut de biologie physico-chimique
3 LitCh - Lipotoxicity and Channelopathies - ConicMeds
STIM - Signalisation et Transports Ioniques Membranaires
4 IJM (UMR_7592) - Institut Jacques Monod
5 LBM UMR 7203 - Laboratoire des biomolécules
6 UPC - UFR SDV - Université Paris Cité - UFR Sciences du Vivant [Sciences]

Abstract : Build hydrogen atoms from united-atom molecular dynamics of lipids and calculate the order parameters

Type de document : Logiciel
Domaine : Informatique [cs] / Bio-informatique [q-bio.QM]

Liste complète des métadonnées Voir

CONSULTER

Software Heritage

swh:1:dir:4c63d5ca3497726a1e54ac152ce1667d7c004d2b;origin=https://github.com/patrickfuchs/buildH/;visit=swh:1:snp:a63a8d07dbebeb442a06707be476817cec44ac72;anchor=swh:1:rev:9f05672515e1cdb0064eeb34f63844296193bc0d

Consulter

https://hal.archives-ouvertes.fr/hal-03467422
Contributeur : Pierre Poulain Contacter le contributeur
Soumis le : lundi 6 décembre 2021 - 15:02:04
Dernière modification le : mercredi 20 avril 2022 - 16:16:23

MÉTADONNÉES

Keywords : lipids | order parameters | molecular dynamics simulation | united-atom

version 1.6.0

Licences https://spdx.org/licenses/BSD-3-Clause

Langage de programmation Python

Code Repository https://github.com/patrickfuchs/buildH

COLLECTIONS

ESPCI | STIM | ENSCP | UP-SCIENCES | PSL | LBM | ENS-PARIS | UNIV-POITIERS | PARISTECH | SORBONNE-UNIVERSITE | UNIV-TOURS | SU-SCIENCES | LBT | ESPCI-PSL | IJM | CNRS | INC-CNRS | INSERM | UNIV-PARIS

CITATION

Hubert Santuz, Amélie Bâcle, Pierre Poulain, Patrick Fuchs. buildH. 2019, (swh:1:dir:4c63d5ca3497726a1e54ac152ce1667d7c004d2b;origin=https://github.com/patrickfuchs/buildH/;visit=swh:1:snp:a63a8d07dbebeb442a06707be476817cec44ac72;anchor=swh:1:rev:9f05672515e1cdb0064eeb34f63844296193bc0d). (hal-03467422)

EXPORTER

CodeMeta | BibTeX | TEI | DC | DCterms | EndNote | Datacite

https://hal.inria.fr/hal-03467422

En résumé

Archivez

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

Décrivez avec des métadonnées

`README`, **`LICENSE`**, **`AUTHORS`**, **`codemeta.json`**

Référencez

SWHID plutôt que DOI, contexte

Citez

Version, release, fichier, ligne

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Merci !

The first 5 years of Software Heritage in 5 minutes!

Software Heritage FAQ

Open Science tutorial: source code deposit (SWH + HAL)