Making Software a first class citizen in the scholarly world

Archive - Reference - Describe - Cite (ARDC)

Roberto Di Cosmo
Inria and Université de Paris

SEFM 2020, *Amsterdam*

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

Software source code: a precious part of our heritage

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

"Programs must be written for people to read, and only incidentally for machines to execute."

Apollo 11 source code (excerpt)

```
P63SPOT3    CA      BIT6        # IS THE LR ANTENNA IN POSITION 1 YET
            EXTEND
            RAND    CHAN33
            EXTEND
            BZF     P63SPOT4    # BRANCH IF ANTENNA ALREADY IN POSITION 1

            CAF     CODE500     # ASTRONAUT:    PLEASE CRANK THE
            TC      BANKCALL    #                SILLY THING AROUND
            CADR    GOPERF1
            TCF     GOTOPOOH    # TERMINATE
            TCF     P63SPOT3    # PROCEED        SEE IF HE'S LYING

P63SPOT4    TC      BANKCALL    # ENTER          INITIALIZE LANDING RADAR
            CADR    SETPOS1

            TC      POSTJUMP    # OFF TO SEE THE WIZARD ...
            CADR    BURNBABY
```

Quake III source code (excerpt)

```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    x2 = number * 0.5F;
    y = number;
    i = * ( long * ) &y; // evil floating point bit level hacking
    i = 0x5f3759df - ( i >> 1 ); // what the fuck?
    y = * ( float * ) &i;
    y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
// y = y * ( threehalfs - ( x2 * y * y ) ); // 2nd iteration, this
can be removed

    return y;
}
```

Len Shustek, Computer History Museum

"Source code provides a view into the mind of the designer."

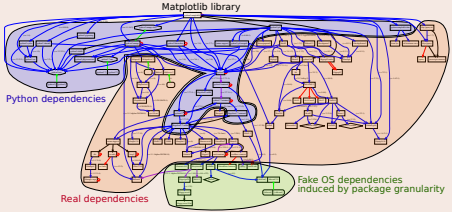
Source code is *special* (software is *not* data)

Software *evolves* over time

- projects may last decades
- the *development history* is key to its *understanding*

Complexity

- *millions* of lines of code
- large *web of dependencies*
  - easy to break, difficult to maintain
  - *research software* a thin top layer
- sophisticated *developer communities*



Precious, endangered *executable* and *human readable* knowledge

key people passing away, platforms (GoogleCode, Gitorious, etc.) closing down …

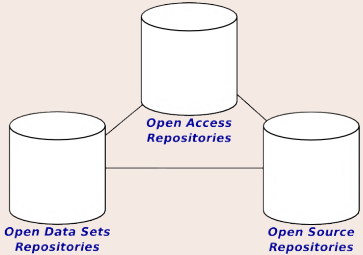
no organised effort to catalog and archive it

Software Source code: a pillar of Open Science

Three pillars of Open Science

Sometimes, if you dont have the software, you dont have the data

C. Borgman, Paris, 2018



N.B.: links are essential

A plurality of needs

Researcher

- archive and reference software used in articles
- find useful software
- get credit for developed software
- verify/reproduce/improve results

Laboratory/team track software contributions

- produce reports / web page

Research Organization know its software assets

- technology transfer
- impact metrics

Where we stand

lack of reproducibility in SE and CS ..

- no replication studies (Zannier et al., ICSE 2006)
- only 20% installable tools in TOSEM 2001 to 2006 (Ghezzi, ICSE 2009)
- 601 mainstream papers: 508 with tools, only 40% *installable* (Collberg, 2015)

main reasons: source code (*or the right version of it*) cannot be found

a recent awakening (~2010)

- Policies: Artifact Evaluation (AEC), ACM Artifact Review and Badging, ...
- Working groups: FORCE11, RDA, SPSO, ...
- Journals: IPOL, ReScience, InsightJournal, JOSS, eLife, ACM DL, ...
- Generalist Repositories: FigShare, Zenodo, (but here software is *just data*)

but a lot is left to be done!

What is at stake: ARDC in increasing order of difficulty

Archive

Research software artifacts must be properly archived

make sure we can *retrieve* them (*reproducibility*)

Reference

Research software artifacts must be properly referenced

make sure we can *identify* them (*reproducibility*)

Describe

Research software artifacts must be properly described

make it easy to *discover* and *reuse* them (*visibility*)

Cite/Credit

Research software artifacts must be properly cited *(not the same as referenced!)*

to give *credit* to authors (*evaluation*!)

We need an infrastructure *designed for* software source code: *now we have one!*

Outline

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

An international, non profit initiative built for the long term

Sharing the vision

United Nations Educational, Scientific and Cultural Organization

And many more ...
www.softwareheritage.org/support/testimonials

Donors, members, sponsors

INVENTEURS DU MONDE NUMÉRIQUE

Platinum sponsors

Gold sponsor

Silver sponsors

Bronze sponsors

… raising awareness about Software Source Code

UNESCO, Inria, Software Heritage invite

40 international experts meet in Paris …

Their call is published on Feb 2019

It's an important *policy tool*, already referenced and used … *yes, you can sign it!*

https://en.unesco.org/foss/paris-call-software-source-code

The largest software archive, a shared infrastructure

Cultural Heritage | **Industry** | **Research** | **Education**

Software Heritage

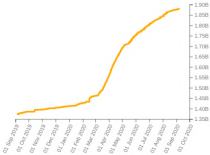
As of today the archive already contains and keeps safe for you the following amount of objects:

Source files

8,846,381,610

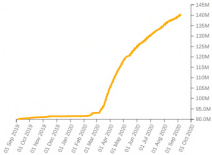
Commits

1,880,663,008



Projects

140,348,311



Directories

7,506,954,410

Authors

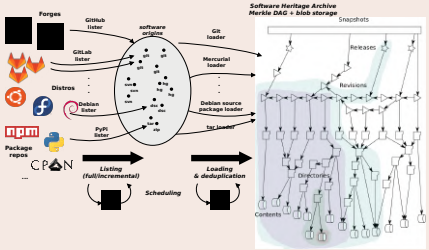
38,603,337

Releases

15,051,940

Addressing the four ARDC needs (see ICMS 2020 for details)

Archive (8B+ files, 140M+ projects)



- save.softwareheritage.org
- deposit.softwareheritage.org

Describe

- *Intrinsic metadata* from source code
- Contributed the Codemeta generator

Reference (20 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs

Now supported in SPDX 2.2, Wikidata etc.

Cite/Credit

- Contributed *software citation* style biblatex-software, v 1.2-2 now on CTAN

Outline

A walkthrough

- Browse the archive
- Get and use SWHIDs (full specification available online)
- Example use in a research article: compare Fig. 1 and conclusions
  - in the 2012 version
  - in the updated version using SWHIDs and Software Heritage
- Cite software with the biblatex-software style from CTAN
- Example use in a research article: extensive use of SWHIDs in a replication experiment
- Trigger archival of your preferred software in a breeze
- Curated deposit in SWH via HAL, see for example: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, ...
- Rescue landmark legacy software, see the SWHAP process with UNESCO

A word on the trust model for systems of identifiers

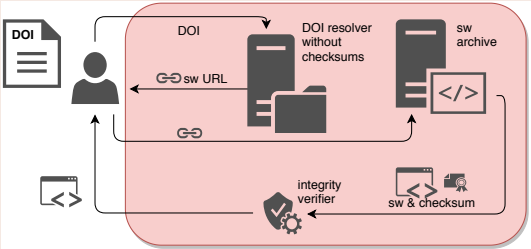
Two general classes of systems of identifiers

intrinsic *computed* from the object *(no registry required, fully decentralised)*
(e.g.: chemical notation, music notation, hashes, SWHIDs)

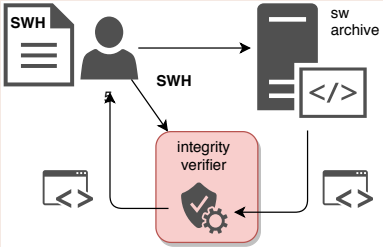
extrinsic *assigned* by an authority *(need a registry)*
(e.g.: passport number, DOI, ARK, RRID, etc.)

See the dedicated blog post for more details

Trust model, extrinsic (e.g. DOIs)



Trust model, intrinsic (e.g. SWHIDs)



Outline

Adoption is coming …

HAL software curated deposit workflow

Curated Archiving of Research Software Artifacts
International Journal of Digical Curation, 2020

Reference archive for swmath.org

swMATH
an information service for mathematical software

See *code* links, e.g. SemiPar package

Image Processing On Line (IPOL)

- archives
- reference
- cite: see BibLaTeX example

JTCAM (Theor. Comp. and Appl. Mech)

- instructions for authors recommend archival in Software Heritage
- biblatex-software in journal LaTeX class

Policy

NATIONAL PLAN
FOR OPEN SCIENCE

now officially in the *French National Plan for Open Science*

Self archival guidelines

Software Heritage
1 Prepare your public repository
README, AUTHORS & LICENSE files
2 Save your code
http://save.softwareheritage.org/
3 Reference your work
(full repository, specific version or code fragment)

- online summary
- full ICMS 2020 paper

Breaking news, and a lesson to be learned

Saving 250.000 endangered repositories…

- summer 2019: BitBucket announce Mercurial VCS phase out
- fall 2019: Software Heritage teams up with Octobus (funded by NLNet, thanks!)
- july 2020: BitBucket erases *250.000* repositories
- august 2020: bitbucket-archive.softwareheritage.org is live

… preserving the web of knowledge (Tweet is here)

Bottomline

explicit deposit is important, …

… and we must promote it…

… but will never be enough.

(think also of all software dependencies!)

Summing up: a revolutionary infrastructure *designed for source code*

Software Heritage

www.softwareheritage.org

global source code *archive* — *Library of Alexandria of source code*

- harvest *all* software, not just research software
- save code now to trigger archival on demand
- API for curated deposit

universal intrinsic identifiers

SWHIDs provide standard independent of version control systems

uniform data model, *full graph* of development history

enables large scale, big code research

Let's all make research software a first class citizen!

- leverage Software Heritage in conferences, journals, AEC for *archival* and *reference*
- adopt and promote `biblatex-software` to *cite* software artifacts
- join the conversation on *software citation* and *software evaluation* criteria
- tackle the scientific problems : big code, classification, infrastructure, etc.

Thank you!

R. Di Cosmo
Archiving and Referencing Source Code with Software Heritage
ICMS 2020 (https://dx.doi.org/10.1007/978-3-030-52200-1_36)

R. Di Cosmo, M. Gruenpeter, S. Zacchiroli
Referencing Source Code Artifacts: a Separate Concern in Software Citation,
CiSE 2020 (10.1109/MCSE.2019.2963148) (hal-02446202)

P. Alliez, R. Di Cosmo, B. Guedj, A. Girault, M.-S. Hacid, A. Legrand and N. Rougier
Attributing and referencing (research) software: Best practices and outlook from Inria,
CiSE 2020 (10.1109/MCSE.2019.2949413) (hal-02135891)

J.F. Abramatic, R. Di Cosmo, S. Zacchiroli
Building the Universal Archive of Source Code, CACM, October 2018 (10.1145/3183558)

Appendix

Outline

Software Heritage for Research and Innovation

Reference platform for *Big Code*

- unique observatory of all software development
- big data, machine learning paradise: classification, trends, coding patterns, code completion…

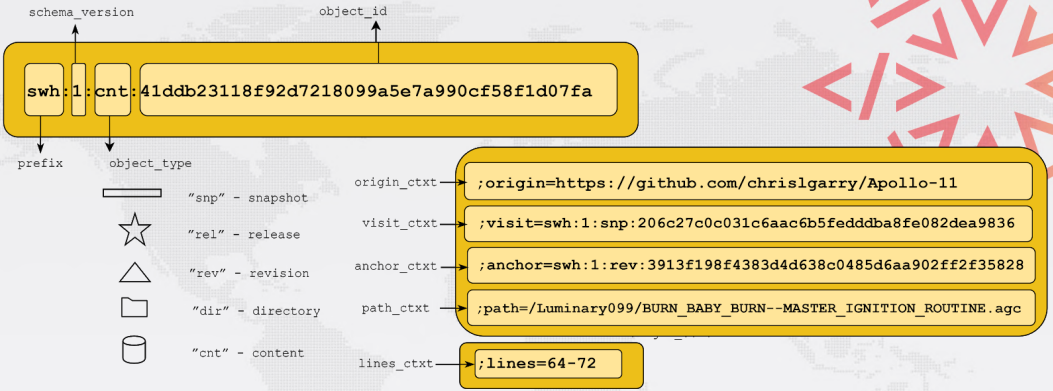
First datasets are available!

- full graph of software development (~20Bn nodes, ~200Bn edges) see Pietri, Spinellis, Zacchiroli, MSR 2019 https://dx.doi.org/10.1109/MSR.2019.00030
- MSR 2020 mining competition see https://2020.msrconf.org/track/msr-2020-mining-challenge#Call-for-Papers

Outline

Software Heritage Identifiers (SWHID)

link to full docs



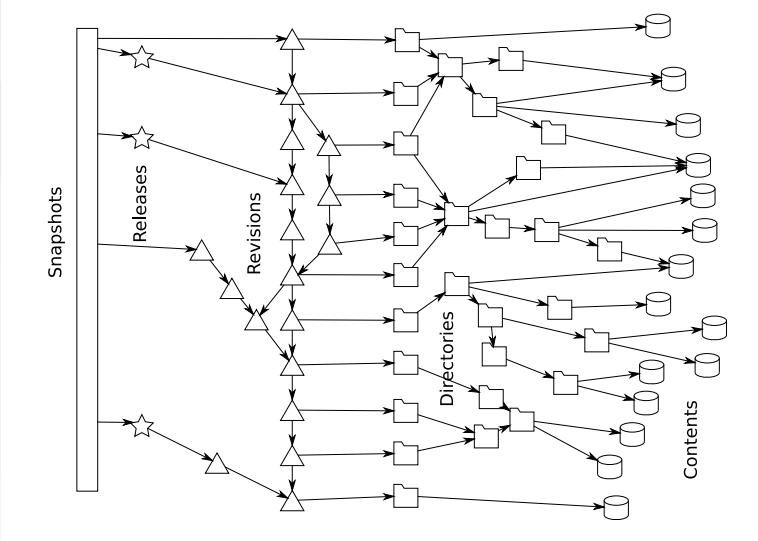
An emerging standard

- in Linux Foundation's SPDX 2.2
- IANA registered, WikiData property P6138

Examples:

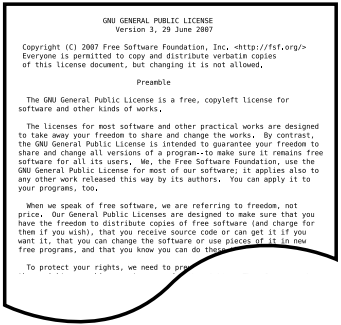
- Apollo 11 AGC excerpt,
- Quake III rsqrt

A worked example



A worked example

Contents



GNU GENERAL PUBLIC LICENSE
Version 3, 29 June 2007

Copyright (C) 2007 Free Software Foundation, Inc. <http://fsf.org/>
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

Preamble

The GNU General Public License is a free, copyleft license for software and other kinds of works.

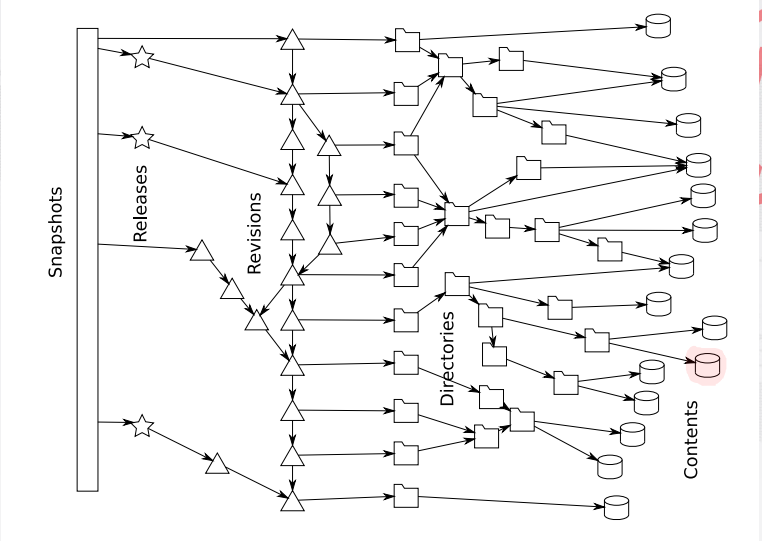
The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change all versions of a program--to make sure it remains free software for all its users. We, the Free Software Foundation, use the GNU General Public License for most of our software; it applies also to any other work released this way by its authors. You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for them if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs, and that you know you can do these

To protect your rights, we need to pre

sha1: 8624bcdae55baeef...
sha256: 8ceb4b9ee5aded...
sha1_git: 94a9ed024d385...
length: 35147

A worked example



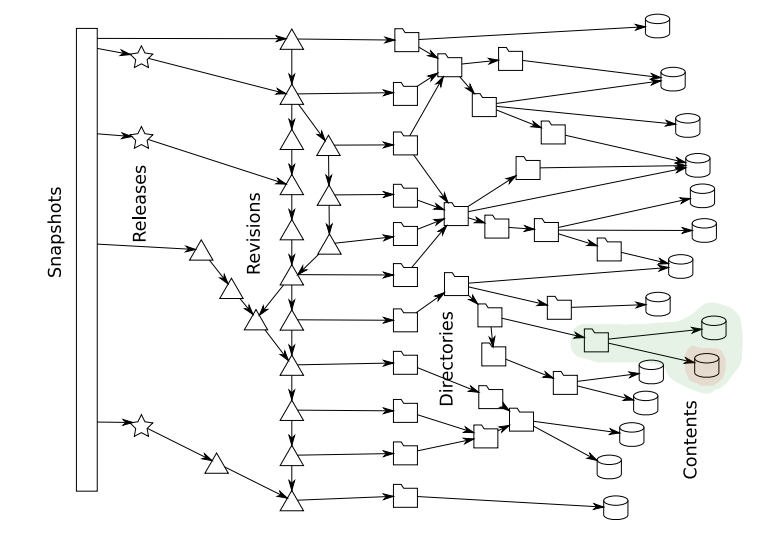
A worked example

Directories

```
100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d0a34af6f52cf3cf6b0c2f7bd0648fbd255e77f AUTHORS
100644 blob 94a9ed024d3859793618152ea559a168bbcbb5e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e0867751dfb095c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5eaf2eb4 Makefile
100644 blob 5c7e3a5bbddb038682ba7793f440492ed9678bb3 Makefile.local
100644 blob 8617980629cd24e6080404f09aa749b085b3e07b README.db_testing
100644 blob 76b29f94cf815e0869c414d38d78d7ce08ec514e README.dev
040000 tree e1e10ecef948af0b93adb0372afc89f12e92618a bin
040000 tree 83e56d0beaf7793c77a45a345c80fcb8af503013 debian
040000 tree a34c9c4ba213f0cedc67f9816348d27955577af5 docs
100644 blob f2a6d32c6135aa7287bbd76167b01df2ae4f1539 requirements.txt
100755 blob eee147c36caf1bbc2d820da8dc026cb5b68180bc setup.py
040000 tree 224bb4c1f4c67fca1d160bffd2d06094e7e1abf3 sql
040000 tree 8631c9cd77bbe993168107ab5baf51f40c6300be swh
040000 tree 8fb905b56ba8ed692f1209b2773b474c6c1d66c1 utils
```

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d

A worked example



A worked example

Revisions

Details | Changes | Files

SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6

Author: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)

Committer: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)

Subject: **provenance.tasks: add the revision -> origin cache task**

Parent: fc3a8b59ca1df424d860f2c29ab07fee4dc35d10 : *test_storage: properly pipeline origin and cont...*

provenance.tasks: add the revision -> origin cache task

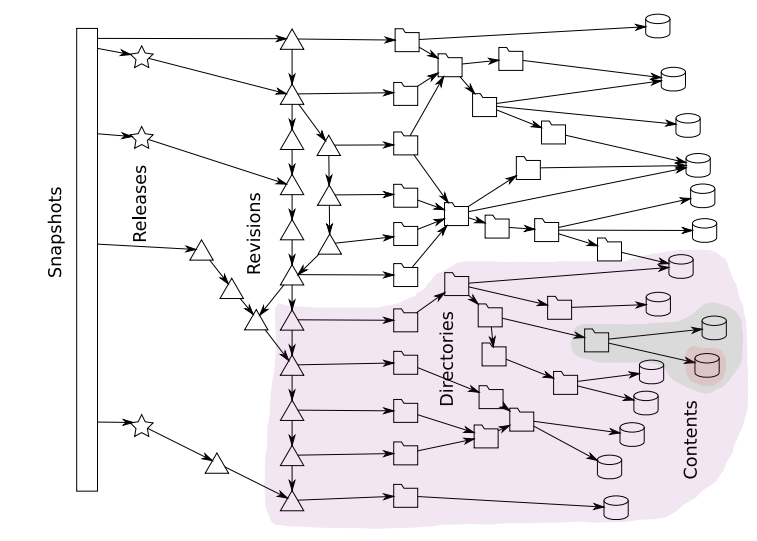
swh/storage/provenance/tasks.py 77

tree 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
parent fc3a8b59ca1df424d860f2c29ab07fee4dc35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

provenance.tasks: add the revision -> origin cache task

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6

A worked example



A worked example

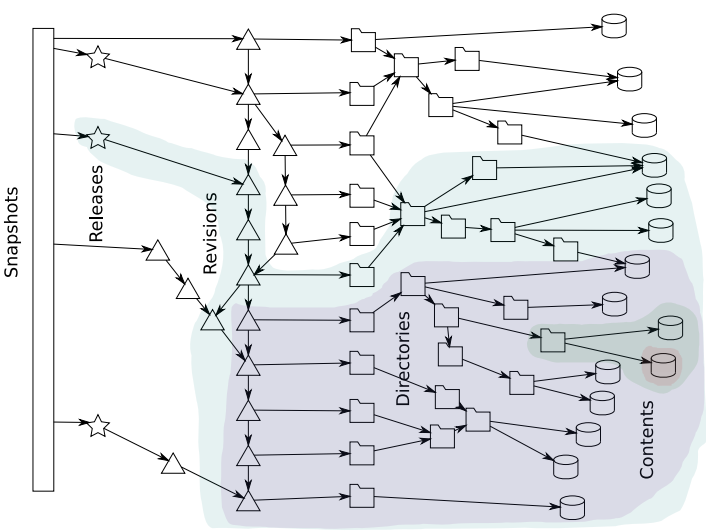
Releases

```
tag v0.0.51
Tagger: Nicolas Dandrimont <nicolas@dandrimont.eu>
Date:   Wed Aug 24 14:36:03 2016 +0200

Release swh.storage v0.0.51

 - Add new metadata column to origin_visit
 - Update swh-add-directory script for updated API
[...]

commit c0c9f16b1e134f593e7567570a1761b156e6eb1d
```

```
object c0c9f16b1e134f593e7567570a1761b156e6eb1d
type commit
tag v0.0.51
tagger Nicolas Dandrimont <nicolas@dandrimont.eu> 1472042163 +0200

Release swh.storage v0.0.51

 - Add new metadata column to origin_visit
 - Update swh-add-directory script for updated API
-----BEGIN PGP SIGNATURE-----

iQIzBAABCAAdBQJXvZTNFhxuaWNvbGFzQGRhbmRyaW1vbnQuZXUACgkQ7AWLMo2+
neqorw//aq6SOb5DijzEa+kWN3rXgVS+1K1vEVh1wNKAwx8eKJ7aX2kEiLDtt7uf
ahpZ6pz3q8nqs6aC1+YrxBfcih3L2YtrdZeWXWqr8xWNMaEoYDb8qaphwh8AD5t2
ICBIiZujbXuCrDt93eKKPwvzZXg+hB0sMWy35Dr6jW7Z7K4Mu/PGglylHPY55yo
IGEndWno7VfH1Vm6t1n5qB7I5mXRaqA+becqddubTZ2xjj+jplUqC8cyqN3hm/fL
qsj2mu8kyz3t8tG/H1/pV+I5OwBlnPoS5TH0tujojEVgPK/dHSP79QuHDHZFkCao
kIj6kAWyU80Mxb+nKV/jeLbrR3+yWBFj3Qp5a1/V8oOTh6E1dALcN6MpEaKCoKtMt
d/gMRaxlI1/g0EDfnsW67G6sDwKPKPHngfVLQ3nV3GaQQTnu1RpMzO06H9/tAwzC
Gg/K1PdHT4hzOil46wYPZyje0U2VXGFu6vVU9vFQ4ZR/WJn+02kMzdcRdrIJSUOMn
RpTTFUsbXUeXHGOpkgXhSYTnvp1gdPc76USTsK0aGe84AZm1Ik0mGrwXCVfPqlYo
nhhibBSHBNMoqyF6yTSOpUbYK70tpYRRUGKWDeRK0wKSxkWKUZGtKzy6JYqIjo29
gulwgZQif5qWQCB0OontAL2+HvPFaVyckMejUhg62cP/+EHIvUk=
=kOxP
-----END PGP SIGNATURE-----
```

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb

A worked example



A worked example

git show-refs

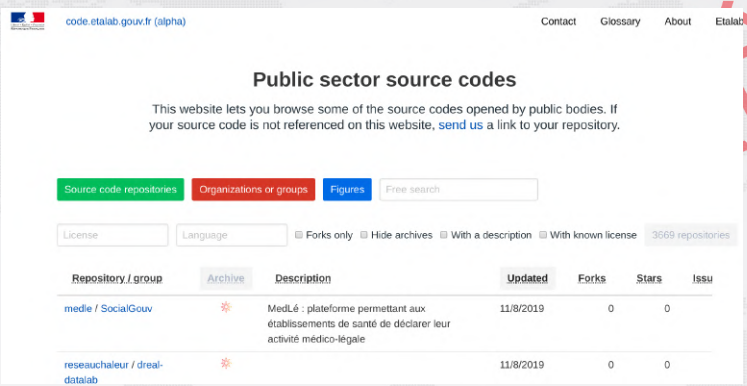
Snapshots

```
commit 08ffeb25770109525eb3ce21691466c53a1d9158 refs/heads/atime
commit ba5443a24e3f9fe323a46c292cec4fcbe61c67eb refs/heads/directory-listing-arrays
commit d69e0dbf892383ff6589b27fbe1c05d27238d9c5 refs/heads/foo
commit cf7ff9eea0eb22f8946908f5a8019f67de468e08 refs/heads/master
commit 7eca197fc66d2024047e54b1ed9e8b44361a0fc2 refs/heads/tmp-directory-add
commit 642a205f37de85005a85d427b53ee4fb2252e82e refs/heads/tmp/generic-releases
tag 20f043b1379cf768d966597799fd4907c757f755 refs/tags/v0.0.1
tag 72a21991a384e539996dbb867bfb0bee72aee2cd refs/tags/v0.0.10
tag 3590e0ca0eb070e5b376705fa230bbfa4ffa5cc refs/tags/v0.0.11
tag 33378427a403ba569a67777b8d58f6674fbc6556 refs/tags/v0.0.12
tag 06f74652755b327cf590311c2bfa036cf3b4b35d refs/tags/v0.0.13
tag 5a6325fe86ab854b581d7442667d92a11e32f3bd refs/tags/v0.0.14
tag 586fba4e580b4f5fab05f599367643cbcb1a9c7f refs/tags/v0.0.15
tag 8cd8b885f4098bf363177742bd289f660e5be51c refs/tags/v0.0.16
tag a542444ee3f0fbed35efb202fee035c809abc7d6 refs/tags/v0.0.17
tag 228a2f1650dd12222e556559462e1e06fc4993d9 refs/tags/v0.0.18
tag 606979a4ca05d497fc0d24aad00dce82636ef47c refs/tags/v0.0.19
tag 32bf5a59fc2a323baa6d5f15a6ad5382ec275a67 refs/tags/v0.0.2
tag 3147c3d31ec46cf6492f881e908b1237ebdff2c7 refs/tags/v0.0.20
tag 215ea50daba111e082e0b72e76eb4b6073a87908 refs/tags/v0.0.21
tag 3fb168c2072a5d6252124257a1e5dfc0f5ffa1df refs/tags/v0.0.22
tag 8cdbee8da4d73fc5d262789e460a16ac3c72aba4 refs/tags/v0.0.23
...
```

id: b464cad1b66fff266a37b46ea6e7a04b545e904b

Outline

News : archiving *public* code



https://code.etalab.gouv.fr

News : SWHAP

Paris Call on Software Source Code

"[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive"

SWHAP : an important step forward

- detailed guidelines to curate landmark legacy source code and archive it on Software Heritage
- intense cooperation with Università di Pisa and UNESCO
- open to all, we'll promote it worldwide

https://www.softwareheritage.org/swhap

News : ENEA mirror

Thomas Jefferson, February 18, 1791

…let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.

Welcoming ENEA

Italian National Agency for New Technologies, Energy and Sustainable Economic Development

- first institutional mirror
- increased resilience
- AI infrastructure for researchers
- stepping stone to an European joint effort