Software Heritage

a revolutionary infrastructure for Open Source

Roberto Di Cosmo
Director, Software Heritage
Inria and Université Paris Cité

October 2024

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Introduction
2. Software Heritage
3. What can Software Heritage do for OSPOs?
4. Conclusion

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- *35+ years* of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- *25+ years* of Free and Open Source Software
- *15+ years* building and directing structures for the common good

1999 *DemoLinux* – first live GNU/Linux distro

2007 *Free Software Thematic Group*
150 members 40 projects 200Me

2008 *Mancoosi project* www.mancoosi.org

2010 *IRILL* www.irill.org

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

2021 *EOSC Task Force on Infrastructures for Software*, European Union

Outline

1. Introduction
2. Software Heritage
3. What can Software Heritage do for OSPOs?
4. Conclusion

Unveiled in 2016

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 and share all software source code

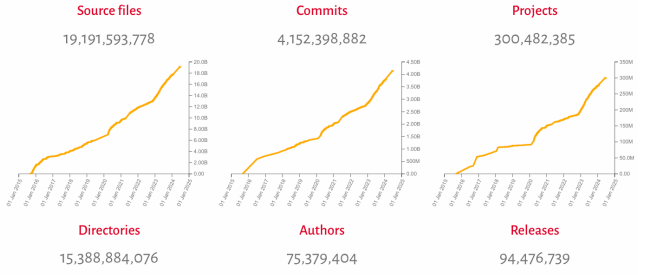
Research infrastructure

enable analysis of all software source code

A *universal* software archive, as a shared infrastructure

One infrastructure
open and shared

The largest archive ever built



Bitbucket 2,509,402 origins	56,983 origins	git 24,600 origins
26,599 origins	debian 136,338 origins	53,297 origins
GitHub 197,883,004 origins	gitiles 10,171 origins	GitLab 4,216,298 origins
git 2,926 origins	Gogs 172 origins	GO 971,549 origins
Guix 14,482 origins	GNU 354 origins	heptapod 1,207 origins
launchpad 503,631 origins	Maven 312,461 origins	NixOS 14,482 origins

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

Diamond sponsor

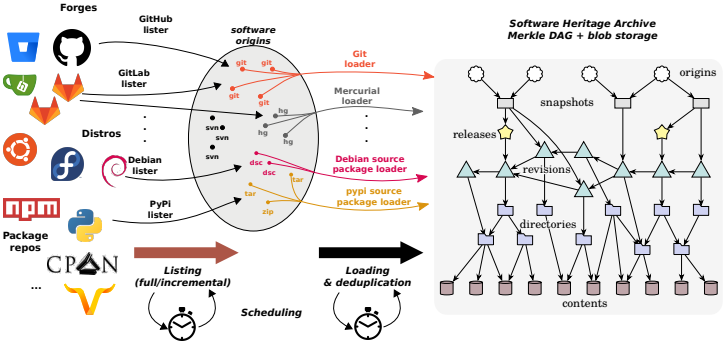
Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors

The archive under the hood



Global development history permanently archived in a uniform data model

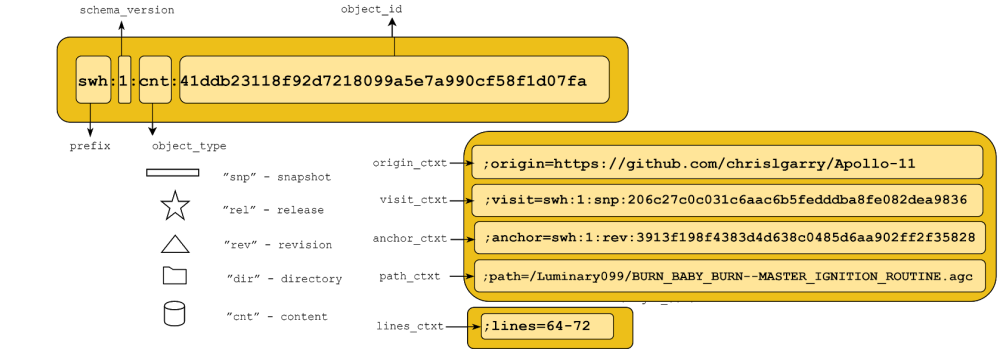
- over 20 billion unique source files from over 300 million software projects
- ~2PB (compressed) blobs, ~50 B nodes, ~700 B edges

The Software Hash persistent identifier (SWHID)

Software Hash Identifiers (SWHID)

see swhid.org

50+B intrinsic, decentralised, cryptographically strong identifiers, SWHIDs



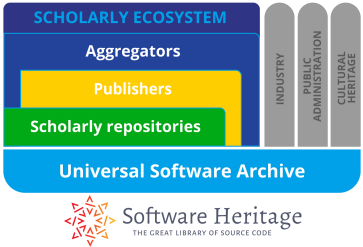
Examples: Apollo 11 AGC, Quake III rsqrt

In SPDX 2.2; IANA registered "swh:"; WikiData P6138

Standardisation ongoing DIS 18670

Software Heritage: key infrastructure for Open Science

EOSC SIRS report: Software Source Code and Open Science, 2020



Connect scholarly ecosystem with the whole software ecosystem

See e.g. the French public administration open source catalog

Ongoing work: FAIRCORE4EOSC

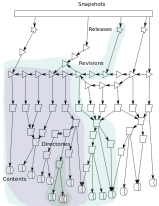
A full workpackage:

- connectors with InvenioRDM (Zenodo), episcience, Dagstuhl, swMath, etc.
- Software Heritage mirror for the European Open Science Cloud (EOSC)
- standardisation of CodeMeta and SWHID

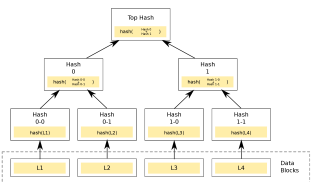
A revolutionary infrastructure

Modern "Library of Alexandria", *international, non profit, long term* initiative
addressing the needs of *industry, research, culture and society as a whole*

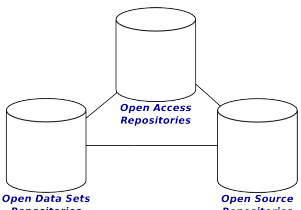
Software Graph



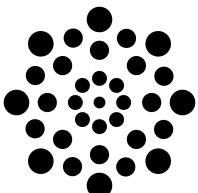
Software Blockchain



Open Science pillar



Big Code



One infrastructure, *shared*: more efficient, less waste …

… supporting revolutionary products!

Outline

1. Introduction
2. Software Heritage
3. What can Software Heritage do for OSPOs?
4. Conclusion

An example is worth 1000 words: the HAL+SWH workflow

Software metadata: codemeta.json

- example from Parmap, created using the Codemeta generator

Integration with the HAL national french open access archive

- Curated deposit: metadata quality due to moderation
- examples: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, …
- export of citation information for biblatex-software
- generation of reports, cv, web pages: for Inria, for CNRS, for CNES, for LIRMM or for Rémi Gribonval using HalTools

Software Heritage + a *curated* metadata repository allows to address all needs …

- *researcher, engineer*: archival, reference, credit, CV etc. *with a little effort from them*
- *labs, organizations*: track and report software production, *curated* catalog

Unique opportunity: your Institutional Portal!

Monitor and showcase your institution's software production

DALL-E's view of an instutional portal

- Curated metadata under institutional authority
- Persistent reference
- Uniform citation
- Automated extraction of reports and indicators

Why SWH?

- platform agnostic, metadata from multiple institutions
- benefit from Software Heritage's future developments

Get involved in the portal specification!

- curation workflow
  researcher initiated (swhid deposit) vs institution initiated (metadata deposit)
- product deployment: on premise vs SaaS
- design of reports and extraction formats

Outline

1. Introduction
2. Software Heritage
3. What can Software Heritage do for OSPOs?
4. Conclusion

A unique opportunity https://softwareheritage.org

Software Heritage is

- vendor neutral
- open source
- a worldwide initiative
- a long term initiative

Software Heritage enables

- long term archival
- reference for reproducibility
- a global software knowledge base
- mutualisation of cost

Call to action

- support Software Heritage as shared open infrastructure, join the ALIG group
- get involved with the new applications

Annual report 2023

5 years in 5 minutes

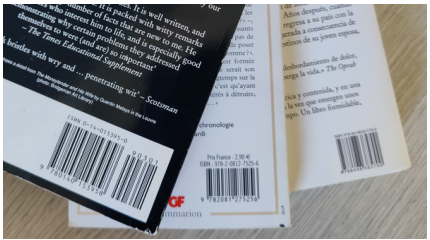
Outline

5 Background on key concepts on identifiers

6 END

Identification vs Location — an example is worth a thousand words

Identification of a book



Goal: identify a book

- one ISBN number per published book
- ISO 2108 standard specification

Location of (a copy of) a book

Goal: find (a copy of) a book

- many locations (locations can change!)
- many approaches for call numbers

we are interested in identification, not in location

Extrinsic vs Intrinsic identifiers

In a nutshell (for more info see this dedicated blog post)

Main difference: how the *relation* between *identifier* and *designated object* is created and maintained. *Persistence* is a key desired property.

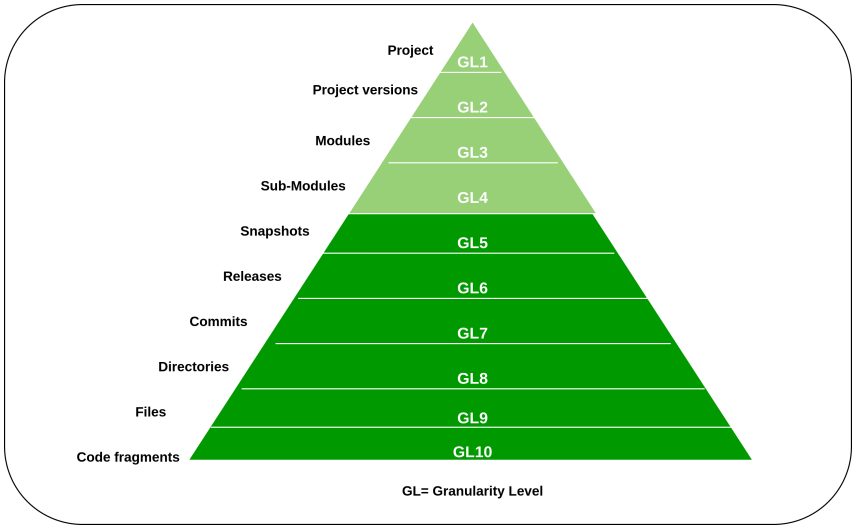
	Extrinsic	**Intrinsic**
relation	register	convention
persistence	external[1]	internal
pre-internet	passport number, ISBN, SSN, etc.	Music/Chemistry notations *e.g. NaCl is table salt*
internet era	DOI, Handle, Ark, etc.	cryptographic hashes *e.g.: git, bitcoin, SWHID*

Software development has *massively adopted* intrinsic identifiers *since 2006*
Git, GitHub, GitLab, pull requests, Guix, Nix, etc. *all rely on* intrinsic identifiers
replicability/reproducibility need intrinsic identifiers
SWHID is *directly compatible* with git!

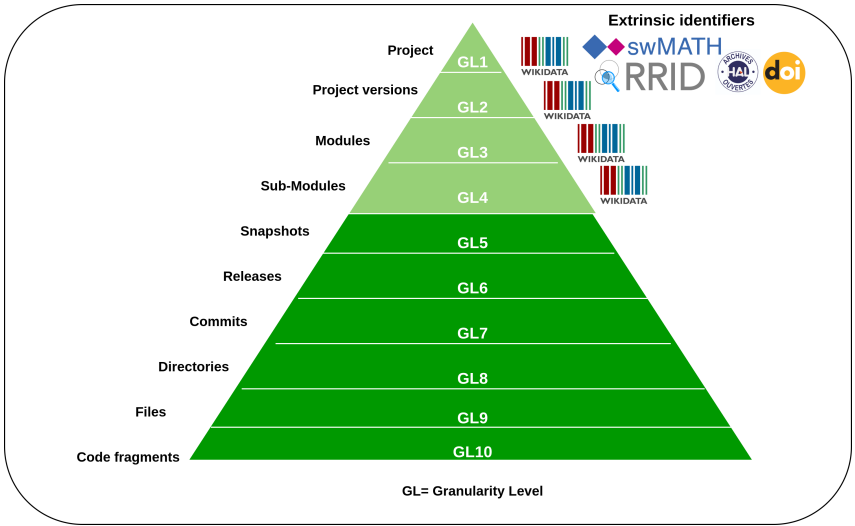
[1] "persistence... is a function of *administrative* care" RFC 3650 (Handle System Overview, 2003)

Choosing identifiers: a matter of granularity

10.15497/RDA00053



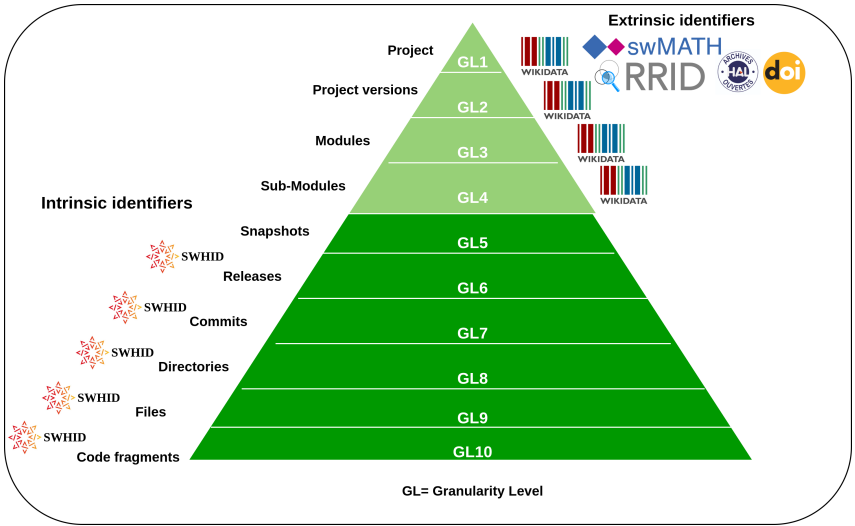
Top concept layers vs. bottom artifact layers



Extrinsic identifiers are key for the concept layers

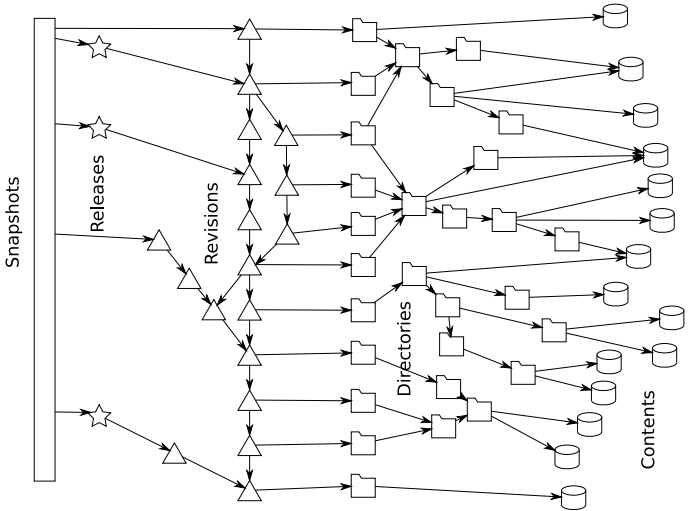
Choosing identifiers: a matter of granularity

10.15497/RDA00053



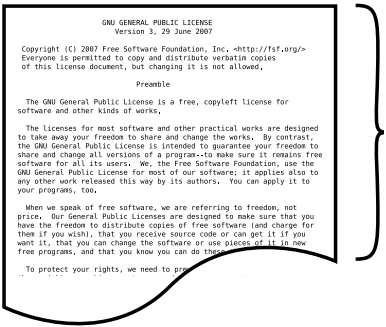
Intrinsic identifiers are key for the artifact layers

SWHID computation: a worked example



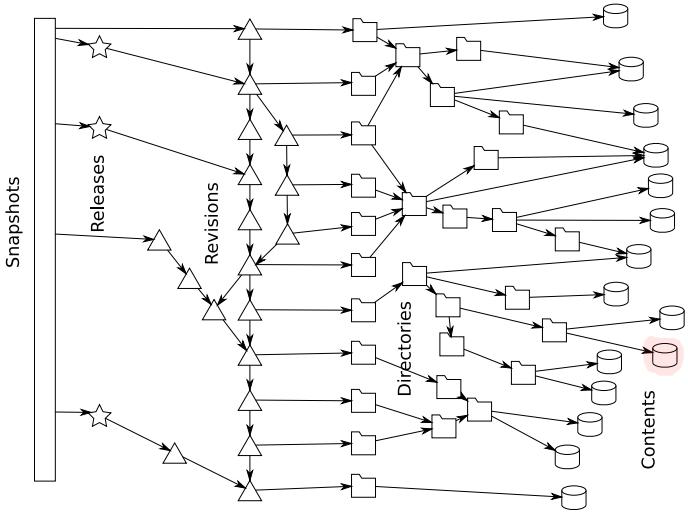
SWHID computation: a worked example

Contents



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

SWHID computation: a worked example



SWHID computation: a worked example

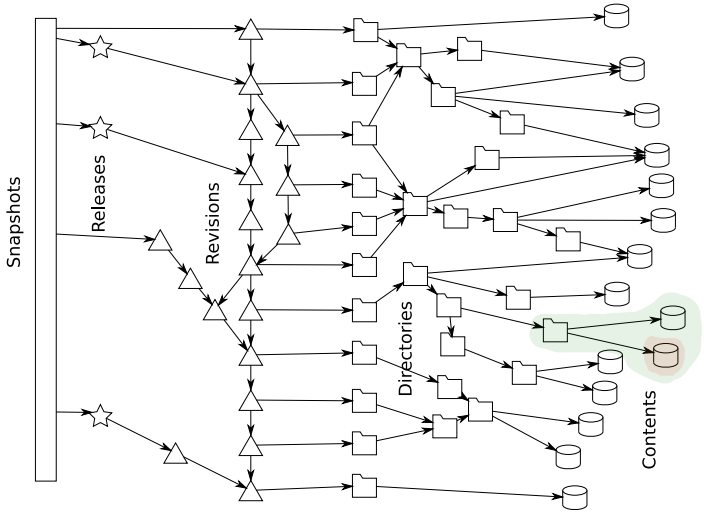
SWHID computation: a worked example

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
```

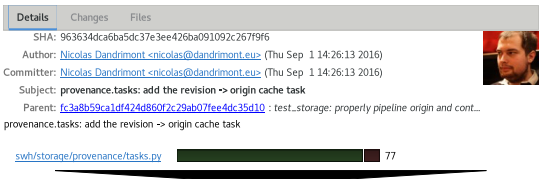
id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d

SWHID computation: a worked example



SWHID computation: a worked example

Revisions

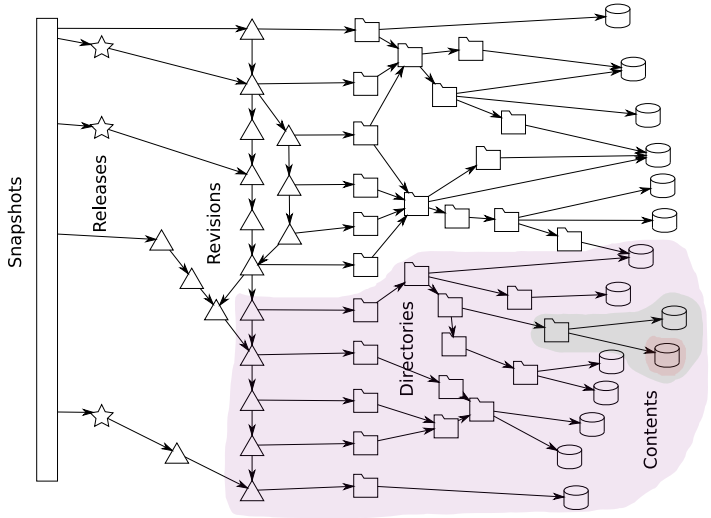


tree 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
parent fc3a8b59ca1df424d860f2c29ab07fee4dc35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

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

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6

SWHID computation: a worked example



SWHID computation: a worked example

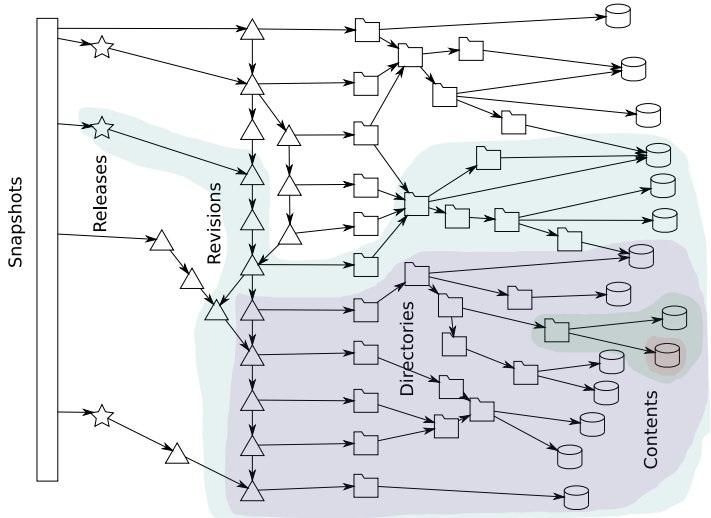
Releases

```
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+kWN3rXqVS+1K1vEVh1wNKAwx8eKJ7aX2kEiLDtt7uf
ahpZ6pz3q8nqs6aC1+YrxBfcih3L2YtrdZeWXWqr8xWNMaEoYDb8qaphwh8AD5t2
ICBlit2ujtXuCrDt93eKKPwvzZXg+hB0sMWy35Dr6jW7Z7K4Mu/PGglyIHPY55yo
lGEndWno7VfH1Vm6t1n5qB7l5mXRaqA+becqddubTZ2xjj+jpIUqC8cyqN3hm/fL
qsj2mu8kyz3t8tG/H1/pV+I5OwBlnPoS5TH0tujoJEvgPK/dHSP79QuHDHZFkCao
klJ6kAWyU80Mxb+nKV/joLbrR3+yWBFj3Op5a1/V8oOTh6E1dALcNMpEaKCoKtMt
d/gMRax1I1/g0EDfnsW67G6sDwKPKPHhgfVLQ3nV3GaQQTnu1RpMzO06H9/tAwzC
Gg/K1PdHT4hzOil46wYPZyje0U2VXGFu6vVU9vFQ4ZR/Wjn+0zMzdcRdrIJSUOMn
RpTTfUsbXUeXHGOpkgXhSYTnvp1gdPc76USTsK0aGe84AZm1Ik0mGrwXCVfPqlYo
nhhibBSHBNMoqyF6yTSOpUbYk70tpYRRUGKWDeRK0wKSxkWKUZGtKzy6jYqIjo29
gulwgZQif5qWQCB0OontAL2+HvPFaVyckMejUhg62cP/+EHlvUk=
=kOxP
-----END PGP SIGNATURE-----
```

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb

SWHID computation: a worked example



SWHID computation: a worked example

git show-refs

Snapshots

```
commit 08ffeb2577010952 5eb3ce21691466c53a1d9158 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 3590e0ca0ebb070e5b376705fa230bbfa4ffa5cc 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 586fba4e580b4f5fab05f59936764cbcb1a9c7f 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