

Archiving, referencing and attributing research software towards software as a first class citizen

Roberto Di Cosmo
Inria and Université de Paris

February 13th, 2020



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

- 1 Software Source Code: a (forgotten) pillar of Science
- 2 Meet Software Heritage
- 3 Archive and *reference all the source code*
- 4 Describe and cite *research source code*
- 5 The road ahead



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)

```
P63SP0T3      CA      BIT6          # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND      CHAN33
              EXTEND
              BZF       P63SP0T4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF      CODE500       # ASTRONAUT: PLEASE CRANK THE
              TC       BANKCALL      # SILLY THING AROUND
              CADR     GOPERF1
              TCF      GOTOP00H      # TERMINATE
              TCF      P63SP0T3      # PROCEED SEE IF HE'S LYING

P63SP0T4      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.”

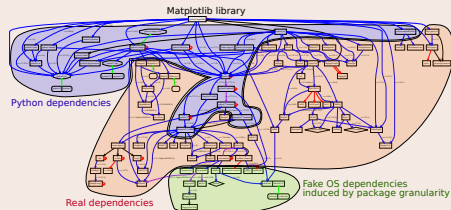
Source code is a *special* and endangered heritage

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
- sophisticated *developer communities*



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

key people **are passing away** ...

no organised effort to catalog and archive it

Software Source code: pillar of Open Science

Software is everywhere in modern research



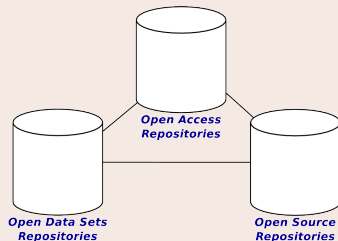
[...] software [...] essential in their fields.

Top 100 papers (Nature, 2014)

Sometimes, if you don't have the software, you don't have the data

Christine Borgman, Paris, 2018

Open Science: three pillars



Nota bene

The links in the picture are **essential**

The state of the art (in CS!) is far from ideal

ICSE (Zannier, Melrik, Maurer, 2006)

- complete absence of replication studies

ACM TOSEM 2001 to 2006

C. Ghezzi <http://bit.ly/tosemreprod>

- 60% of all papers have tools: **only 20% installable**

Collberg's 2015 study

<http://reproducibility.cs.arizona.edu/>

- 601 mainstream papers: 508 with tools, **only 40% installable**

Main reasons

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

A wealth of initiatives!

- Policies: ACM [Artifact Review and Badging](#), ...
- Working groups: [FORCE11](#), [RDA](#), [SPSO](#), ...
- Metrics: [Open Science Monitor](#) (Elsevier!), ...
- Journals: [IPOL](#), [ReScience](#), [InsightJournal](#), [eLife](#), [ACM DL](#), ...
- Repositories: [FigShare](#), [Zenodo](#), ...

but ...

Lack of recognition

not (yet) a first class citizen

- in the EOSC plan
- in the scholarly works

Lack of proper guidance on how to

- *archive* and *reference* software
- choose a license
- *cite* a software project

Archival

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

Identification

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

Metadata

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

Citation

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

A plurality of needs

Researcher

- archive and reference sw used in articles
- get credit for the software they develop
- verify/reproduce/improve results

Laboratory/team

- track software contributions
- produce up-to date report / web page

University/Research Organization

- central view of research software assets
- tech transfer
- impact metrics

- 1 Software Source Code: a (forgotten) pillar of Science
- 2 Meet Software Heritage
- 3 Archive and *reference all the source code*
- 4 Describe and cite *research source code*
- 5 The road ahead





Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share the source code of all the software

Preserving our heritage, enabling better software and better science for all

Reference catalog



find and reference **all** the source code

Universal archive



preserve **all** the source code

Research infrastructure



enable analysis of **all** the source code

Sharing the vision



United Nations
Educational, Scientific and
Cultural Organization



And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors



Platinum sponsors



Gold sponsor



Silver sponsors



Bronze sponsors



The largest software archive, a shared infrastructure

Cultural Heritage



Industry



Research



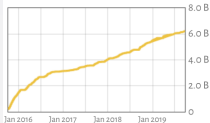
Education



Software Heritage

Source files

6,197,000,081



Commits

1,379,380,527

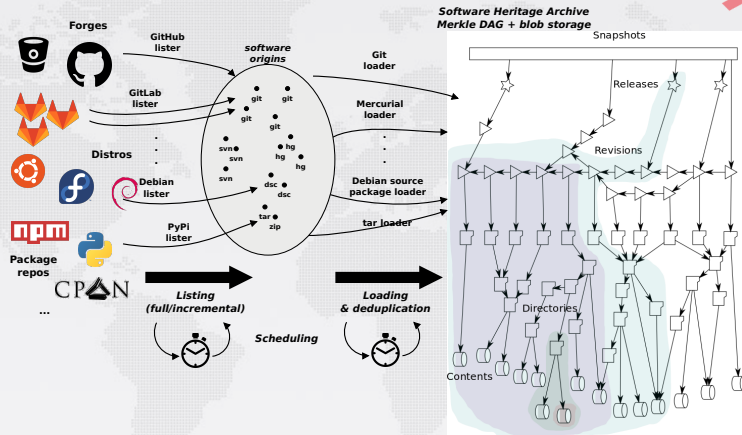


Projects

90,231,104



A peek under the hood



Global development history permanently archived in a *unique* git-like Merkle DAG

- ~400 TB (uncompressed) blobs, ~20 B nodes, ~280 B edges

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>

Raising awareness about Software Source Code

Experts call for greater recognition of software source code as heritage for sustainable development

6 November 2018



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 ...

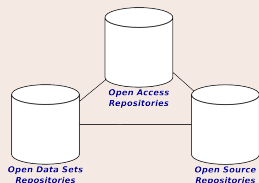
yes, you can sign it!

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

- 1 Software Source Code: a (forgotten) pillar of Science
- 2 Meet Software Heritage
- 3 Archive and reference *all* the source code
- 4 Describe and cite *research* source code
- 5 The road ahead



Software Heritage: a revolutionary infrastructure



- **universal archive** of all source code
 - we archive *all* software: both research and non research
 - we *proactively collect software* in a systematic way
- **intrinsic** identifiers for **reproducibility**
 - identify software artefacts *without any third party*
 - cryptographically strong, compatible with git hashes

Full guidelines available!

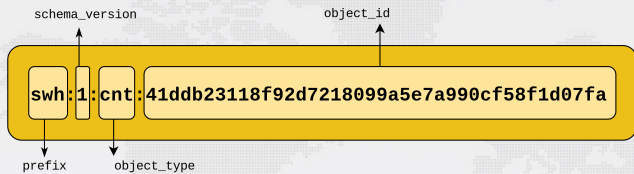
<https://www.softwareheritage.org/save-and-reference-research-software/>

Save code now ... **in just a few clicks**

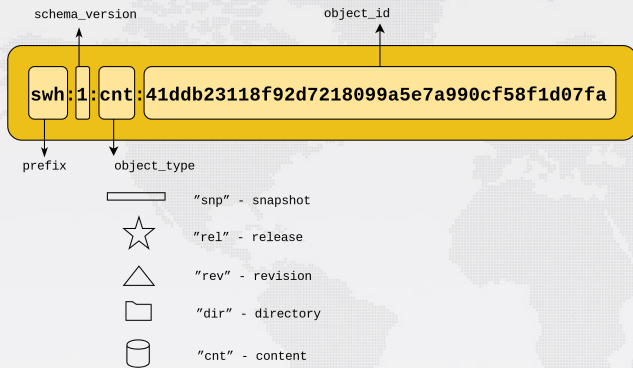
Demo

My 2012 Parmap paper **before** and **after**; other links: **Apollo 11** (and **blog**), **Quake III Arena**

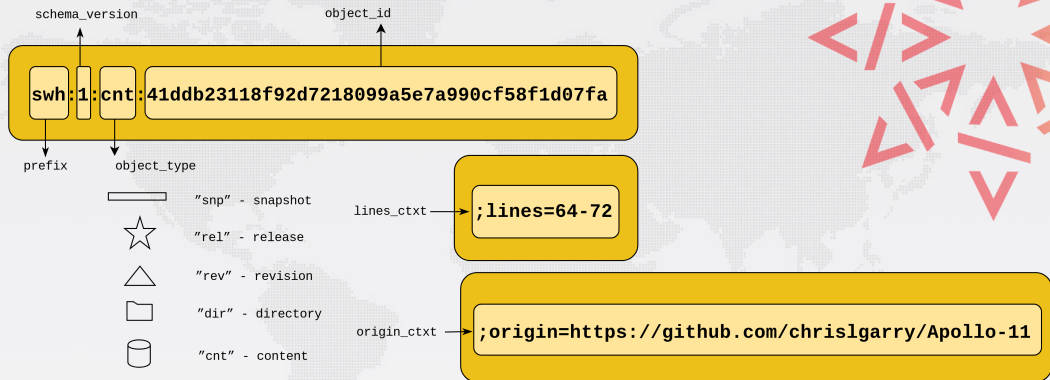
The SWH-ID schema



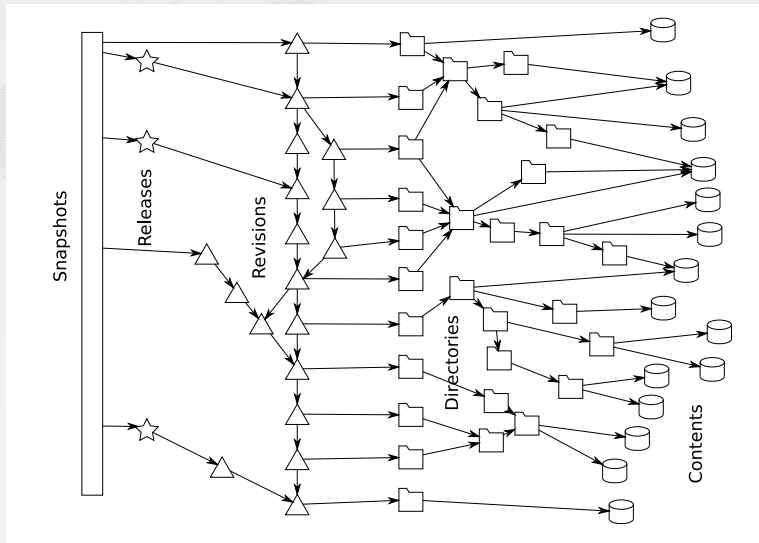
The SWH-ID schema



The SWH-ID schema



A worked example



Contents

```
GNU GENERAL PUBLIC LICENSE
Version 3, 29 June 2007

Copyright (C) 2007 Free Software Foundation, Inc. <https://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.

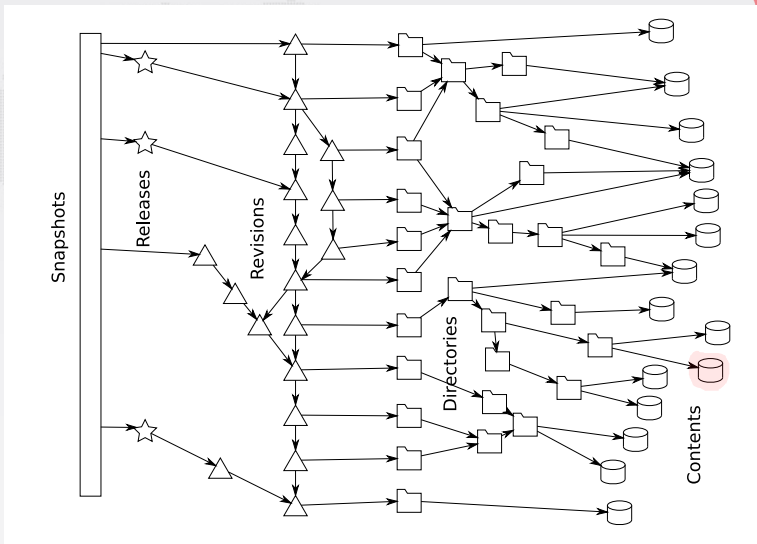
The licenses for most software and other practical works are designed
to take away your freedom to share and change the works. By contrast,
the GNU General Public License is intended to guarantee your freedom to
share and change all versions of a program--to make sure it remains free
software for all 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 things.

To protect your rights, we need to prevent anyone from denying you
```

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

A worked example



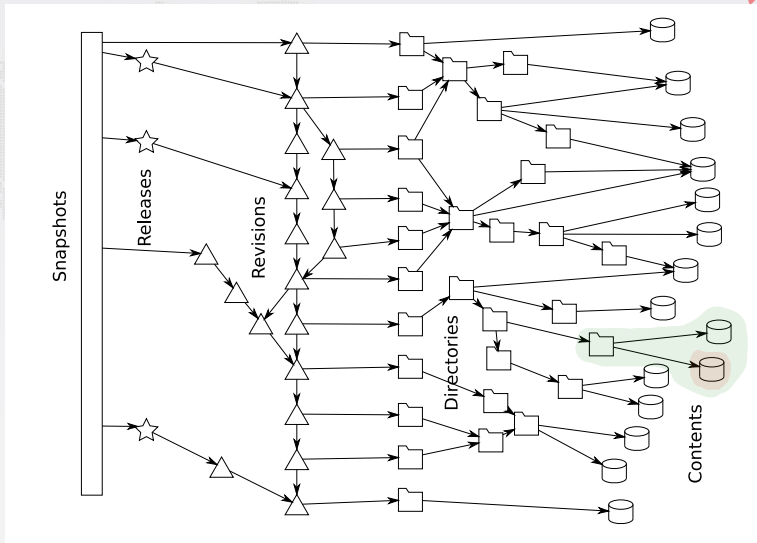


Directories


```
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 e1e10ece948af0b93adb0372afc89f12e92618a bin
040000 tree 83e56d0beaf7793c77a45a345c80fcb8af503013 debian
040000 tree a34c9c4ba213f0cedc67f9816348d27955577af5 docs
100644 blob f2a6d32c6135aa7287bbd76167b01df2ae4f1539 requirements.txt
100755 blob eee147c36caf1bbc2d820da8dc026cb5b68180bc setup.py
040000 tree 224bb4c1f4c67fca1d160bffdd2d06094e7e1abf3 sql
040000 tree 8631c9cd77bbe993168107ab5baf51f40c6300be swl
040000 tree 8fb905b56ba8ed692f1209b2773b474c6c1d66c1 utils
```

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d

A worked example



Revisions

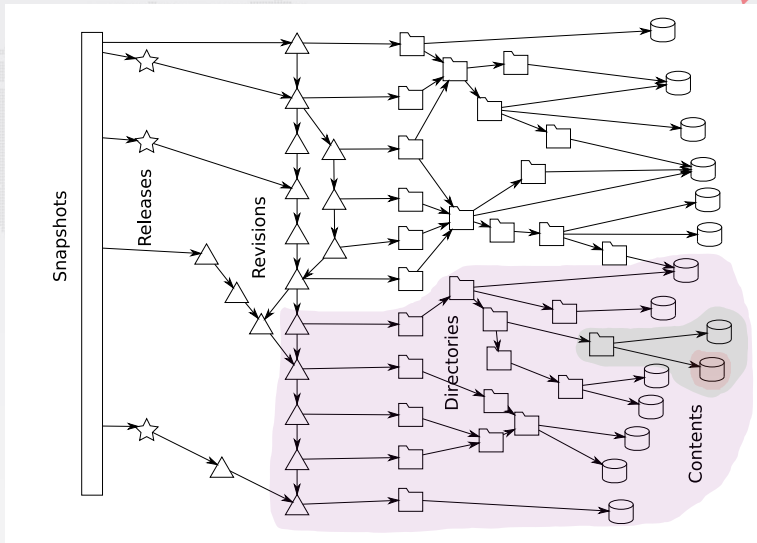
| Details | Changes | Files |
|---|---------|-------|
| SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6 | | |
| Author: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016) | | |
| Committer: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016) | | |
| Subject: provenance.tasks: add the revision -> origin cache task | | |
| Parent: fc3a8b59ca1df424d860f2c29ab07fee4dc35d10 : test...storage: properly pipeline origin and cont... | | |
| provenance.tasks: add the revision -> origin cache task | | |
| swb/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



Releases

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

```
Release sw.h.storage v0.0.51
```

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

```
commit c0c9f16b1e134f593e7567570a1761b156e6eb1d
```

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

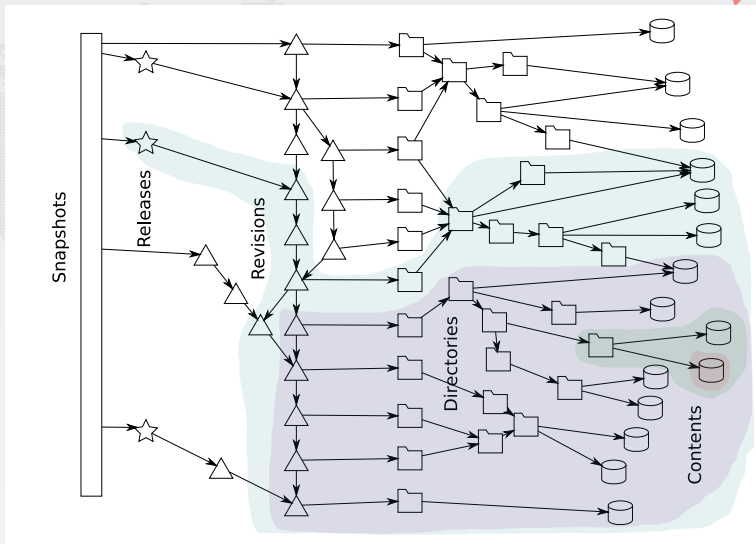
```
Release sw.h.storage v0.0.51
```

```
- Add new metadata column to origin_visit
- Update sw.h-add-directory script for updated API
---BEGIN PGP SIGNATURE---
```

```
iQIzBAABCAAdBQJXvZTNFhXuaWNvbGFzQGRhbmRyaW1vbnQuZlUACgkQ7AWLMo2+
neqorw//aq6SOB5DijzEa+kWN3rXgVS+1K1vEVh1wNKAw8eKJ7aX2kEILDtt7uf
ahpZ6pz3q8nqs6aC1+YrxBfcih3L2YtrdZeWXWqr8xWNMaEoYDb8qaphwh8AD5t2
ICBli2uJXuCrDt93eKKPwvzZXg+h80sMwy35Dr6jW7Z7K4Mu/PgGlyIHPY55yo
IGEndWno7VFH1Vm6t1n5qB7I5mXRaqA+becqddubTZ2xjj+jpUqC8cyqN3hm/fL
qsj2mu8kyz3t8tG/H1/pV+I5OwBlNpoS5TH0tujojEvgPK/dH5P79QuHDHZFkCao
klj6kAWyU80Mxb+nKV/jeLbrR3+yWBFj3Qp5a1/V8o0Th6E1dALcNmPkaKCoKtMt
d/gMRax111/g0EDfnsW67G6sDwKPKPHngfVLQ3nV3GaQQTnu1RpMz006H9tAwzC
Gg/K1PdHT4hz0Jl46wYPZyje0U2VXGFu6vVU9vFQ4ZR/Wjn+0zZdcRdrJlSUOMn
RpTTfUshXUeXHGOpkXhSYTnvp1gdPc76U5TsK0aGe84AZm1Ik0mGrwXCvFPqYo
nhhibB5HBNMoqyF6yTSOpUbyK70tpYRRUGKwDeRk0wK5xkWKUZGtKzy6jYqjJo29
gulwZQif5qWQC80ontAL2+HvFfaVyckMejUhg62cP/+EHlvUk=
=kOxP
---END PGP SIGNATURE---
```

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb

A worked example



Snapshots

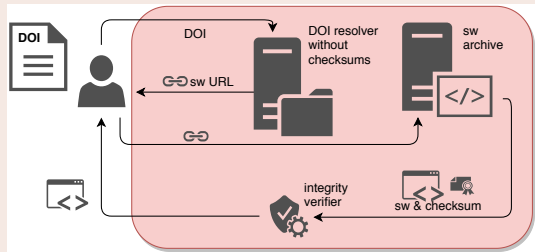
git show-refs

```
commit 08ffeb25770109525eb3ce21691466c53a1d9158 refs/heads/atime
commit ba5443a24e3f9fe323a46c292cec4fcbe61c67eb refs/heads/directory-listing-arrays
commit d69e0dbf892383ff6589b27fbc1c05d27238d9c5 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 586fba4e580b4f5fab05f599367643cbb1a9c7f refs/tags/v0.0.15
tag 8cd8b885f4098bf363177742bd289f660e5be51c refs/tags/v0.0.16
tag a542444ee3f0fbcd35efb202fee035c809abc7d6 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 8cddb0e8da4d731c5d262789e460a16ac3c72aba4 refs/tags/v0.0.23
...
```

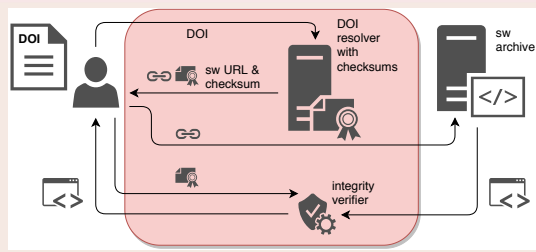
id: b464cad1b66fff266a37b46ea6e7a04b545e904b

Zoom on the trust model for identifiers

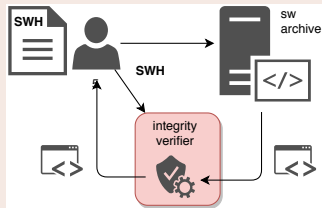
Trust model for usual DOIs



Trust model for DOIs with checksums



Trust model for SWH-IDs



- 
- 1 Software Source Code: a (forgotten) pillar of Science
 - 2 Meet Software Heritage
 - 3 Archive and reference *all* the source code
 - 4 Describe and cite *research* source code
 - 5 The road ahead

Many articles/guidelines

- reproducibility
- archival
- credit and evaluation

Most common limitations

- software is 'just data'
- citation = reference = DOIs
- citation produced by automated tools

A few remarkable exceptions

- ASCL (since 1999): metadata only, carefully curated
- geodynamics.org : source, documentation, metadata
- swmath.org : software catalog via articles

Software Citation WG at Inria (since 10/2018)

- leverage a 50 year experience, make recommendations
- read more <https://hal.archives-ouvertes.fr/hal-02135891>

Why it is not simple

Software is complex

- Structure** monolithic/composite; self-contained/external dependencies
- Lifetime** one-shot/long term
- Community** one man/one team/distributed community
- Authorship** complex set of roles (*more later*)
- Authority** institutions/organizations/communities/single person

Various granularities

Exact status of the source code for reproducibility, e.g.

“you can find at `sw:h:1:cnt:cdf19c4487c43c76f3612557d4dc61f9131790a4;lines=146-187` the core algorithm used in this article”

(Major) release *“This functionality is available in OCaml version 4”*

Project *“Inria has created OCaml and Scikit-Learn”.*

Refined ontology for contributors

- Design, Architecture,
- Coding, Testing, Debugging,
- Documentation, Maintenance, Support,
- Management

see also [CRediT](#), [Geodynamics](#)

Reference is distinct from citation

- **Reference** is for *reproducibility*
- **Citation** is for *credit*

They must not be conflated

Beware of the numbers game:

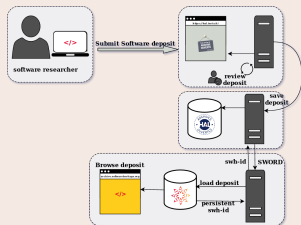
... do we really want an *s-index*?

Keep the human in the loop

When *credit* is at stake, automation/crowdsourcing is not enough!

Humans *are needed* to get *quality information*

How it works, what is special



Generic mechanism:

- SWORD based
- review process
- versioning

Today: deposit .zip or .tar.gz file (*guide*)

Tomorrow: just provide the *SWH id*

Deposit/describe research software in HAL

- author: <https://hal.archives-ouvertes.fr/hal-01872189>
- moderator: <https://hal.archives-ouvertes.fr/hal-01876705>

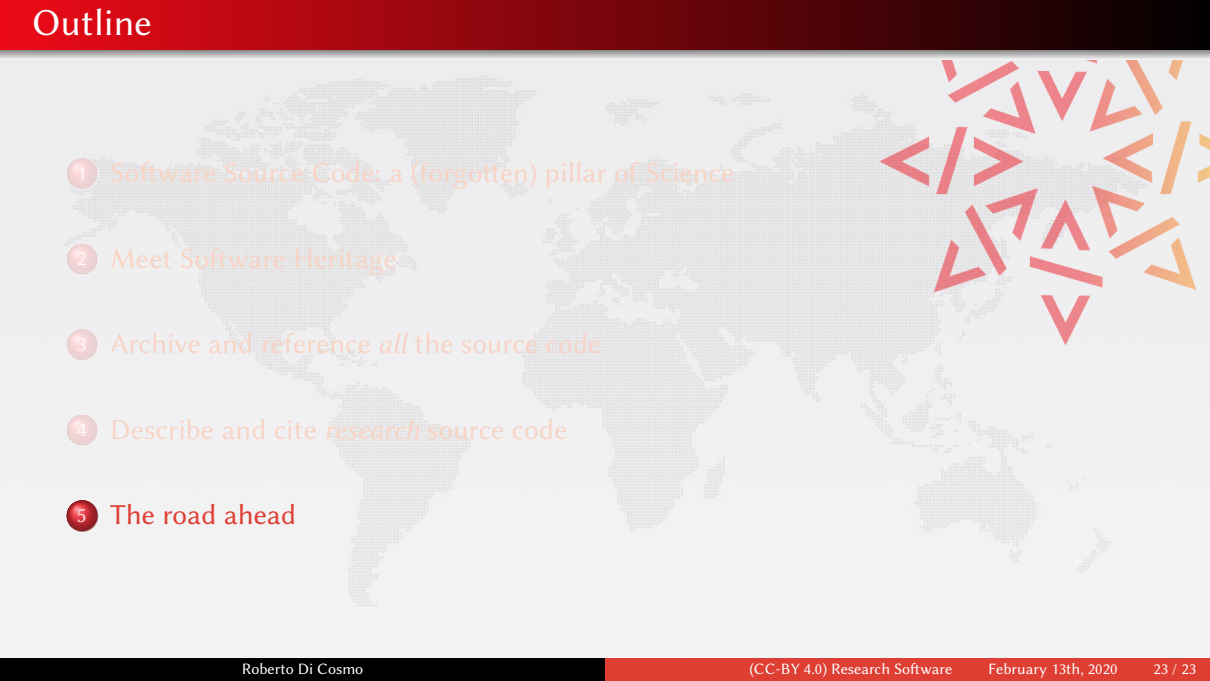
Examples

LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, ...

Article based citation

See for example:

- [SemiPar on swmath.org](#)

- 
- 1 Software Source Code: a (forgotten) pillar of Science
 - 2 Meet Software Heritage
 - 3 Archive and reference *all* the source code
 - 4 Describe and cite *research* source code
 - 5 The road ahead

We need to care more about research software

You can help make a change

- leverage Software Heritage in conferences, journals, AEC for *archival* and *reference*
<https://www.softwareheritage.org/save-and-reference-research-software/>
- join the conversation on *software citation* and *software evaluation* criteria
- tackle the scientific problems : big code, classification, infrastructure, etc.

Thank you!

-  Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchiroli
Building the Universal Archive of Source Code, CACM, October 2018 (10.1145/3183558)
-  Roberto Di Cosmo, Morane Gruenpeter, Stefano Zacchiroli
Referencing Source Code Artifacts: a Separate Concern in Software Citation,
CiSE 2020 (10.1109/MCSE.2019.2963148) (hal-02446202)
-  Pierre Alliez, Roberto Di Cosmo, Benjamin Guedj, Alain Girault, Mohand-Said Hacid, Arnaud Legrand and Nicolas Rougier
Attributing and referencing (research) software: Best practices and outlook from Inria,
CiSE 2020 (10.1109/MCSE.2019.2949413) (hal-02135891)