Software Heritage

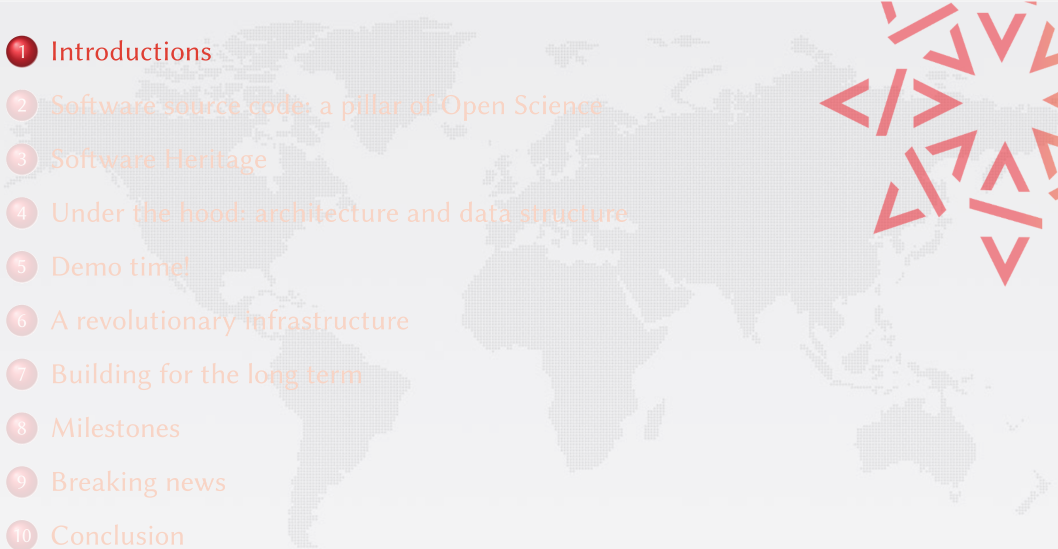
A revolutionary infrastructure for Open Science

Roberto Di Cosmo

October 15th, 2020
La Rochelle

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline



1. Introductions
2. Software source code: a pillar of Open Science
3. Software Heritage
4. Under the hood: architecture and data structure
5. Demo time!
6. A revolutionary infrastructure
7. Building for the long term
8. Milestones
9. Breaking news
10. Conclusion

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

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

1999	*DemoLinux* – first live GNU/Linux distro
2005	*Open Access* debate
2007	*Free Software Thematic Group* 150 members 40 projects 200Me
2015	*Software Heritage* at INRIA
2018	*National Committee for Open Science*, France

Outline

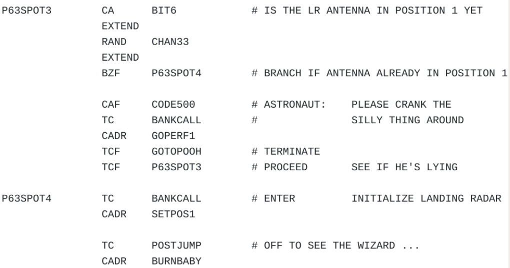
1. Introductions
2. Software source code: a pillar of Open Science
3. Software Heritage
4. Under the hood: architecture and data structure
5. Demo time!
6. A revolutionary infrastructure
7. Building for the long term
8. Milestones
9. Breaking news
10. Conclusion

Source code is *special*

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

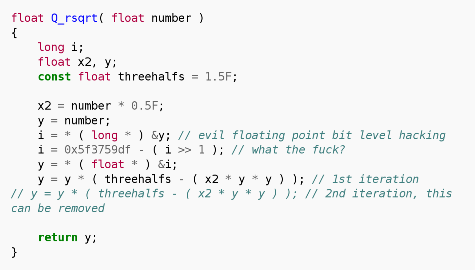
“Programs must be written for people to read, and only incidentally for machines to execute.”

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.”

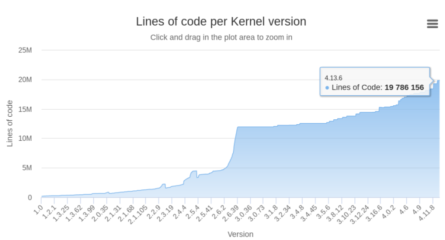
~ 50 years, a lightning fast growth

Apollo 11 Guidance Computer (~60.000 lines), 1969

"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton

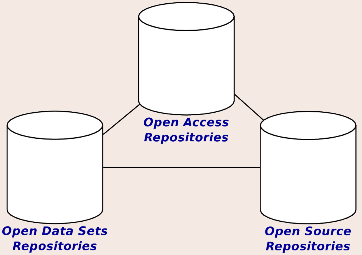
Linux Kernel



… now in your pockets!

Software Source code: pillar of Open Science

Three pillars of Open Science



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

What is at stake in increasing order of difficulty

Archival

Research software artifacts must be properly archived

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

Identification

Research software artifacts must be properly referenced

make 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*!)

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

Outline

1. Introductions
2. Software source code: a pillar of Open Science
3. Software Heritage
4. Under the hood: architecture and data structure
5. Demo time!
6. A revolutionary infrastructure
7. Building for the long term
8. Milestones
9. Breaking news
10. Conclusion

Software Heritage, in a nutshell

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

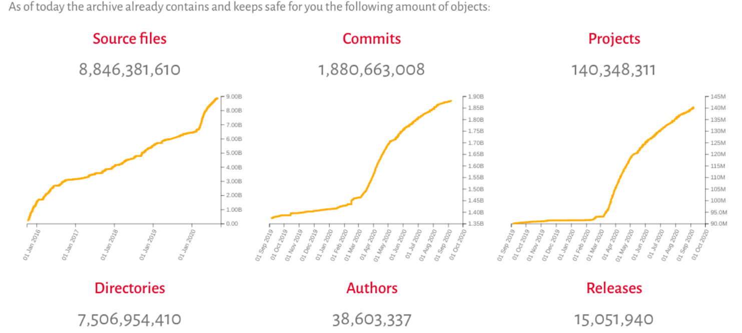
Universal archive

preserve all software source code

Research infrastructure

enable analysis of all software source code

Coverage



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

Source files	Commits	Projects
8,846,381,610	1,880,663,008	140,348,311

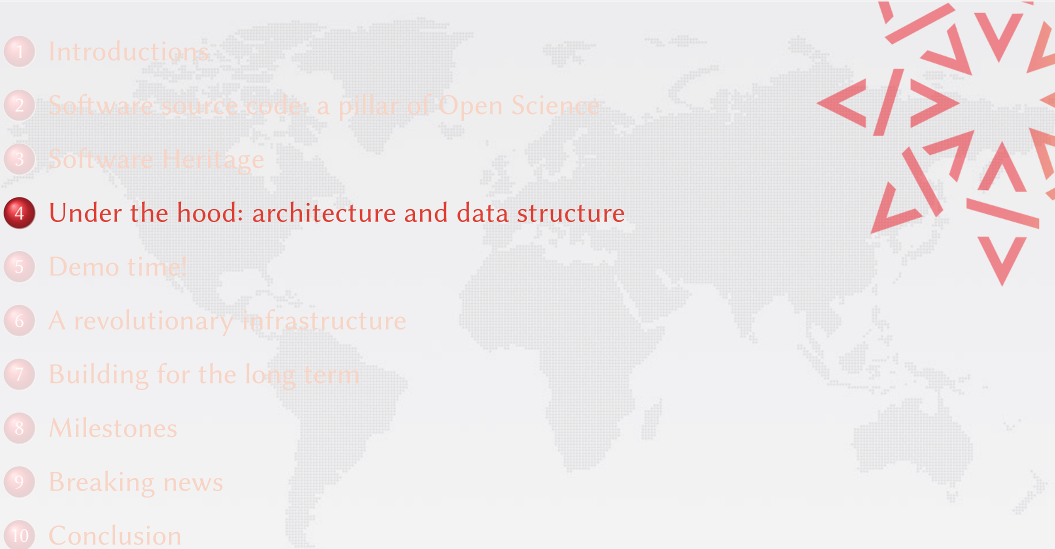
Directories	Authors	Releases
7,506,954,410	38,603,337	15,051,940

GitHub GitLab Bitbucket Google code GITORIOUS Framagit

HAL archives-ouvertes.fr debian npm R GNU Inria inventeurs du monde numérique python Package Index

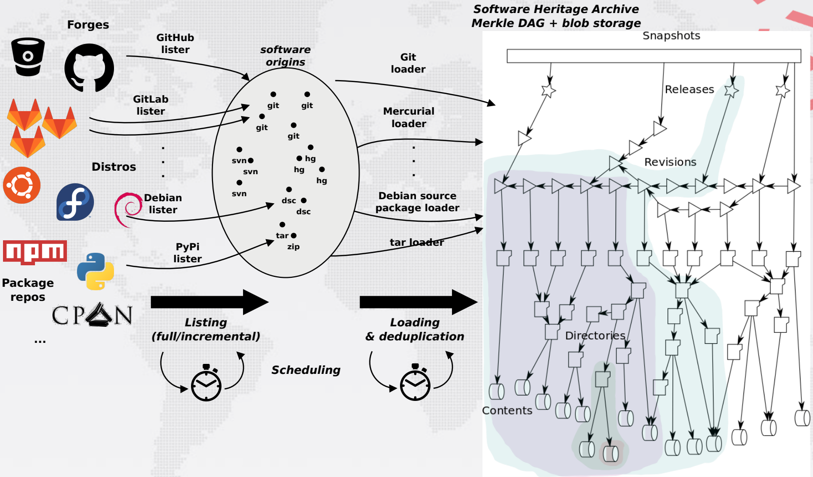
- ~400 TB (uncompressed) blobs, ~20 B nodes, ~300 B edges

Outline



1. Introductions
2. Software source code: a pillar of Open Science
3. Software Heritage
4. Under the hood: architecture and data structure
5. Demo time!
6. A revolutionary infrastructure
7. Building for the long term
8. Milestones
9. Breaking news
10. Conclusion

Automation, and storage



Forges

GitHub lister

GitLab lister

Distros

Debian lister

Package repos

PyPi lister

...

software origins

git git git git svn svn svn svn hg hg hg dsc dsc tar zip

Git loader

Mercurial loader

Debian source package loader

tar loader

Listing (full/incremental)

Loading & deduplication

Scheduling

Software Heritage Archive Merkle DAG + blob storage

Snapshots

Releases

Revisions

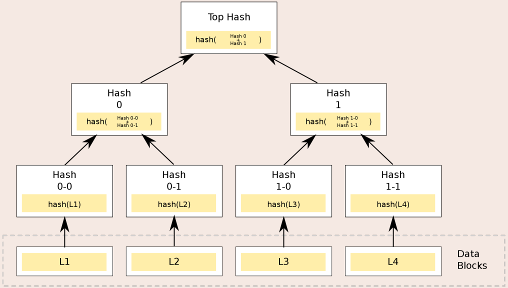
Directories

Contents

Full development history permanently archived in a uniform data model.

Much more than an archive!

Merkle tree (R. C. Merkle, Crypto 1979)



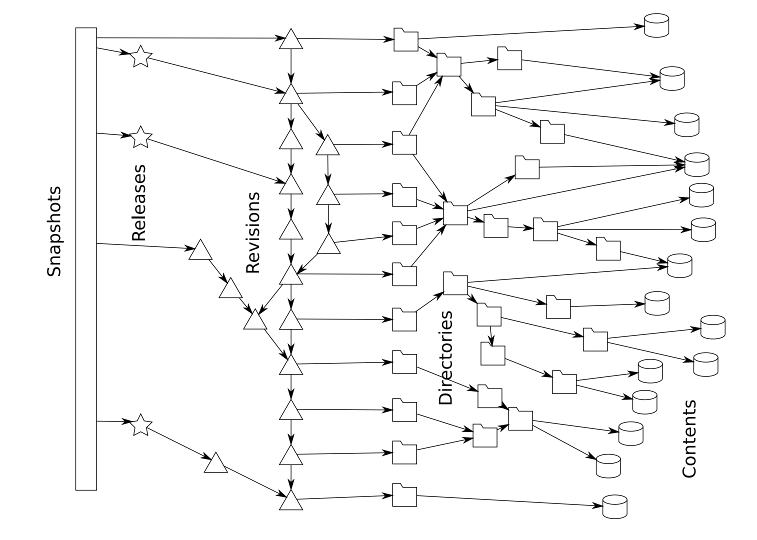
Combination of

- tree
- hash function

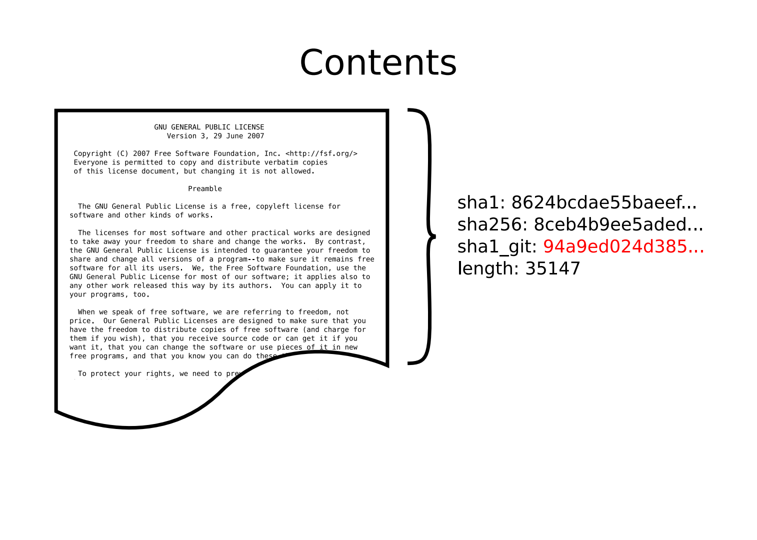
Classical cryptographic construction

- fast, parallel signature of large data structures
- widely used (e.g., Git, blockchains, IPFS, ...)
- built-in deduplication

The archive in pictures



The archive in pictures

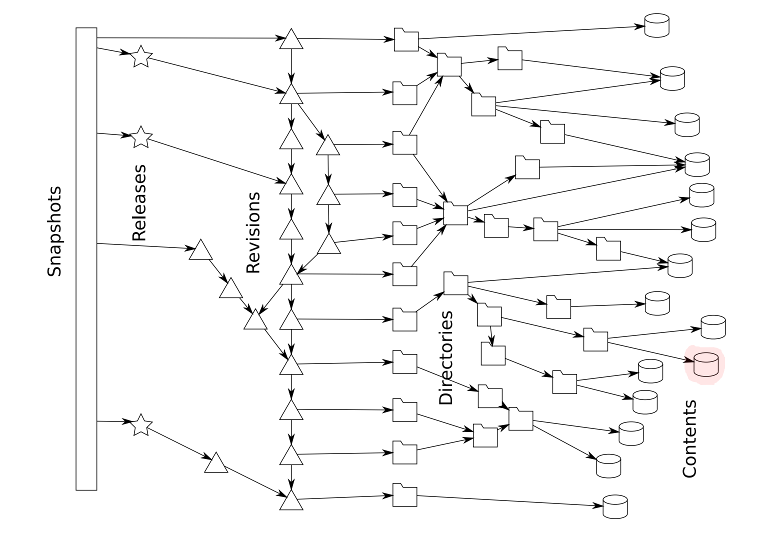


Contents

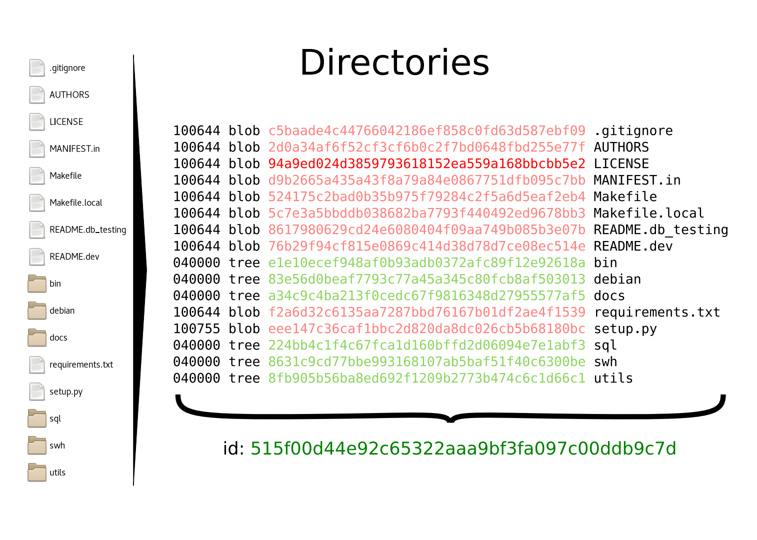
```
                           GNU GENERAL PUBLIC LICENSE
                              Version 3, 29 June 2007

 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

The archive in pictures



The archive in pictures

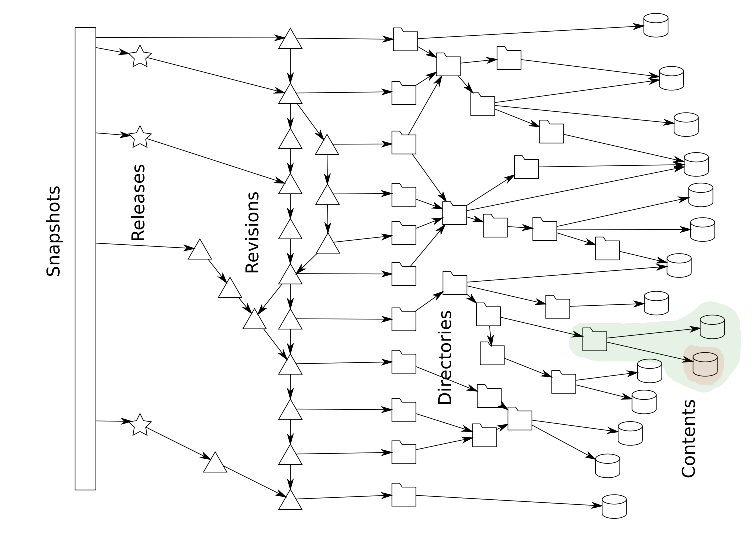


Directories

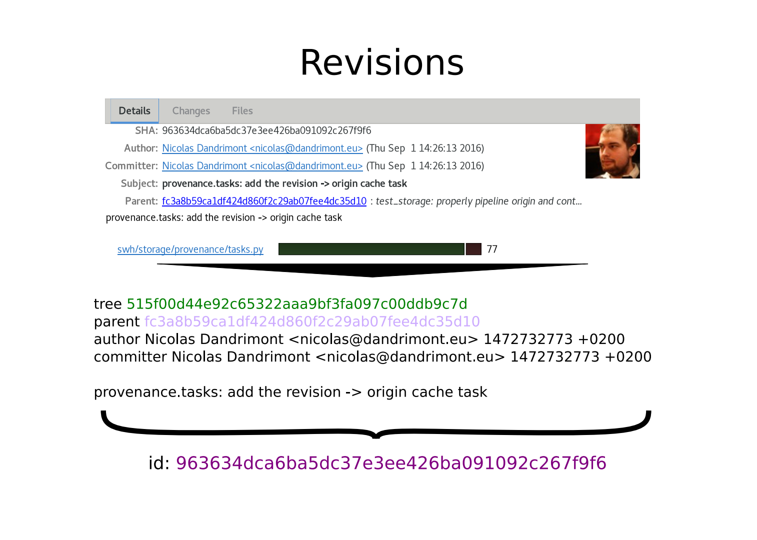
```
100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d0a34af6f52cf3cf6b0c2f7bd0648fbd255e77f AUTHORS
100644 blob 94a9ed024d3859793618152ea559a168bbcbb5e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e0867751dfb095c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5eaf2eb4 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

The archive in pictures



The archive in pictures



Revisions

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...*

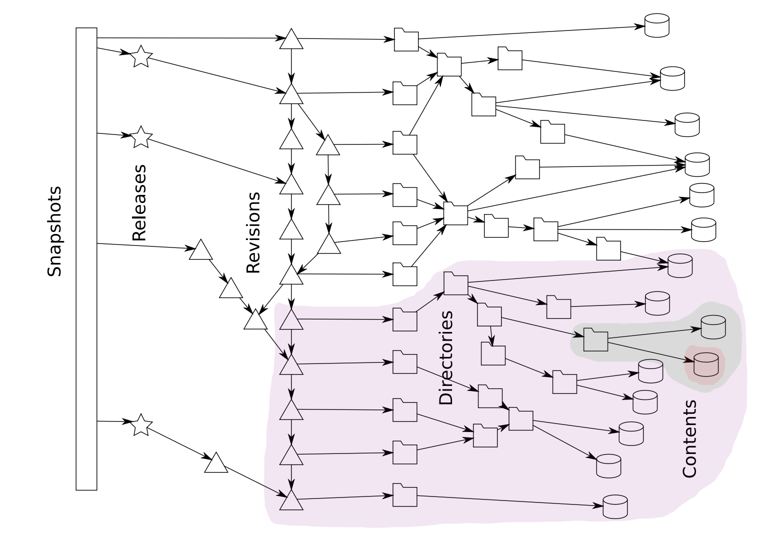
provenance.tasks: add the revision -> origin cache task

swh/storage/provenance/tasks.py 77

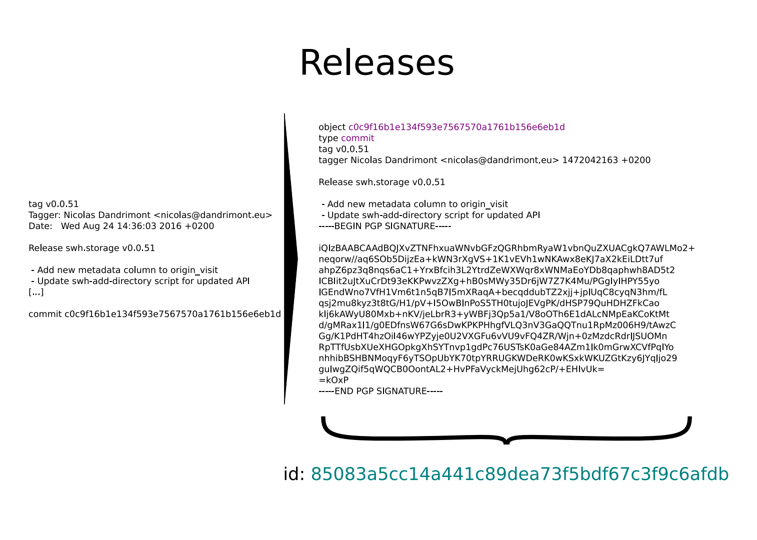
```
tree 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
parent fc3a8b59ca1df424d860f2c29ab07fee4dc35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

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

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6

The archive in pictures



The archive in pictures



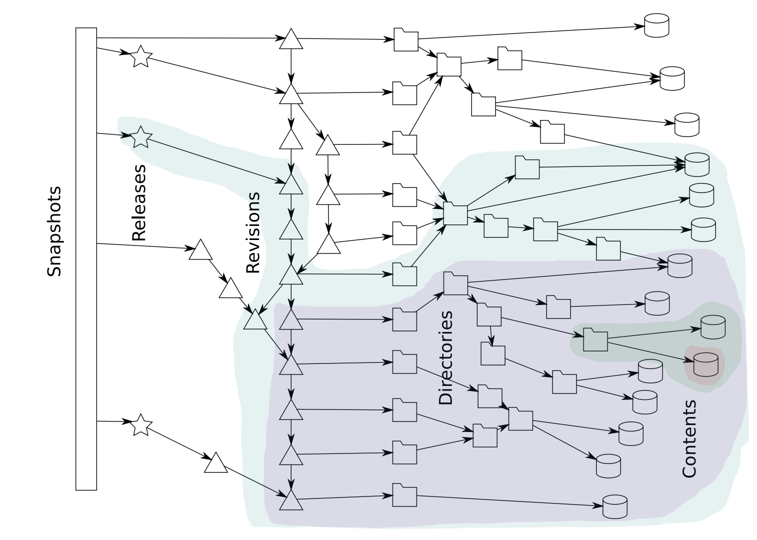
Releases

```
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
```

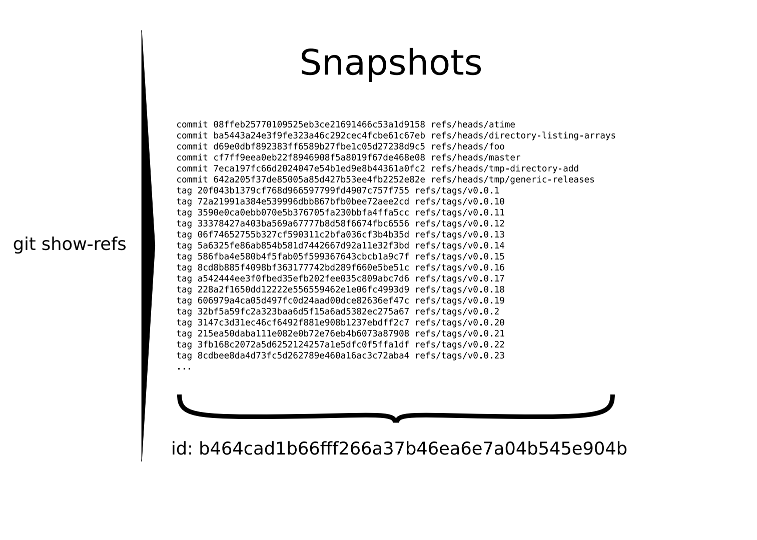
```
object c0c9f16b1e134f593e7567570a1761b156e6eb1d
type commit
tag v0.0.51
tagger Nicolas Dandrimont <nicolas@dandrimont.eu> 1472042163 +0200

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
lCBIit2uJtXuCrDt93eKKPwvzZXg+hB0sMWy35Dr6jW7Z7K4Mu/PGgIyIHPY55yo
lGEndWno7VfH1Vm6t1n5qB7I5mXRaqA+becqddubTZ2xjj+jplUqC8cyqN3hm/fL
qsj2mu8kyz3t8tG/H1/pV+I5OwBlnPoS5TH0tujoJEVgPK/dHSP79QuHDHZFkCao
klj6kAWyU80Mxb+nKV/jeLbrR3+yWBFj3Qp5a1/V8oOTh6E1dALcNMpEaKCoKtMt
d/gMRax1I1/g0EDfnsW67G6sDwKPKPHhgfVLQ3nV3GaQQTnu1RpMz006H9/tAwzC
Gg/K1PdHT4hzOiI46wYPZyje0U2VXGFu6vVU9vFQ4ZR/Wjn+0zMzdcRdrlJSUOMn
RpTTfUsbXUeXHGOpkgXhSYTnvp1gdPc76USTsK0aGe84AZm1lk0mGrwXCVfPqIYo
nhhibBSHBNMoqyF6yTSOpUbYK70tpYRRUGKWDeRK0wKSxkWKUZGtKzy6JYqljo29
gulwgZQif5qWQCB0OontAL2+HvPFaVyckMejUhg62cP/+EHIvUk=
=kOxP
-----END PGP SIGNATURE-----
```

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb

The archive in pictures



The archive in pictures



Snapshots

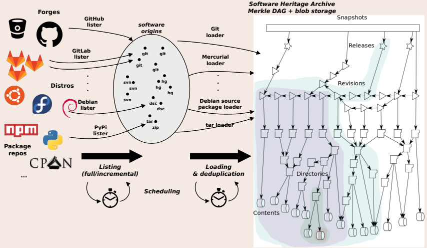
git show-refs

```
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 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 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

Addressing the four needs

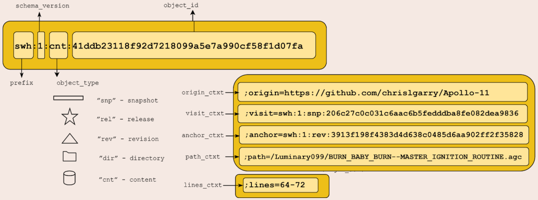
Archive (8B+ files, 130M+ projects!)



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

Reference (20 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs


schema_version
object_id
swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa
prefix
object_type
"snp" - snapshot
"rel" - release
"rev" - revision
"dir" - directory
"cnt" - content
origin_ctxt
;origin=https://github.com/chrislgarry/Apollo-11
visit_ctxt
;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836
anchor_ctxt
;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828
path_ctxt
;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc
lines_ctxt
;lines=64-72

Now supported in SPDX 2.2, Wikidata etc.

Describe

- Intrinsic metadata from source code
- Contributed the Codemeta generator

Cite

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

Outline

1. Introductions
2. Software source code: a pillar of Open Science
3. Software Heritage
4. Under the hood: architecture and data structure
5. Demo time!
6. A revolutionary infrastructure
7. Building for the long term
8. Milestones
9. Breaking news
10. Conclusion

A walkthrough

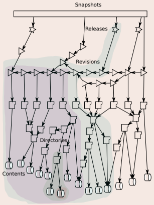
- Browse the archive
- Get and use SWHIDs (full specification available online)
- cite software with the biblatex-software style from CTAN
- Example use in a research article: compare Fig. 1 and conclusions
  - in the 2012 version
  - in the updated version using SWHIDs and Software Heritage
- Example 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

Outline

1. Introductions
2. Software source code: a pillar of Open Science
3. Software Heritage
4. Under the hood: architecture and data structure
5. Demo time!
6. **A revolutionary infrastructure**
7. Building for the long term
8. Milestones
9. Breaking news
10. Conclusion

A revolutionary infrastructure for industry

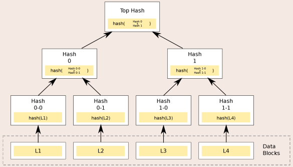
The *graph* of Software Development



All of the software development in a single graph!

- lookup by content hash
- wayback machine for software development
  - `http://archive.softwareheritage.org/`
- … and much more

The *blockchain* of Software Development



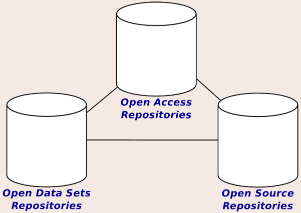
All of a software development… in a single Merkle graph!

Widely used crypto (e.g., Git, blockchains, IPFS, …)

- built-in deduplication
- intrinsic, unforgeable identifiers at all levels
- simplifies traceability (licensing, supply chain management)

A revolutionary infrastructure for research and innovation

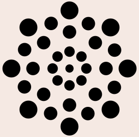
A *pillar* of Open Science



The *reference archive* of Research Software for Open Science

- curated deposit of research software
  - in collaboration with HAL, CCSD and Inria IES
  - now open *to all researchers*!
- intrinsic identifiers for reproducibility

Reference platform for *Big Code*



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

Zoome on big code

First datasets are available!

- full graph of software development (~20Bn nodes, ~200Bn edges) see Pietri et al., MSR 2019 https://dx.doi.org/10.1109/MSR.2019.00030
- MSR 2020 mining competition

Archiving and referencing (your) research software

Save code now

two simple steps to get your repository archived:

- prepare your source code
- go to https://archive.softwareheritage.org/browse/origin/save/

that's it!

Reference code: precisely, and forever!

three simple steps:

- find your code in https://archive.softwareheritage.org
- select your code fragment (optional)
- get the link from the red permalink tab, and use it!

Outline

1. Introductions
2. Software source code: a pillar of Open Science
3. Software Heritage
4. Under the hood: architecture and data structure
5. Demo time!
6. A revolutionary infrastructure
7. **Building for the long term**
8. Milestones
9. Breaking news
10. Conclusion

An international, non profit initiative built for the long term

Sharing the vision

UNESCO
United Nations
Educational, Scientific and
Cultural Organization

IEEE Software · GitLab · ENGINEERING · eclipse
OW2 · fsfe · nlnet FOUNDATION · SIGSOFT · Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE · INAPI
THE LINUX FOUNDATION
ADULLACT · INFORMATICS EUROPE · Computer History Museum · software freedom conservancy
Software Freedom Law Center · AdaCore The GNAT Pro Company · gandi.net · TDF
FREE SOFTWARE FOUNDATION
creative commons · SIF Société Informatique de France · open source initiative · openinventionnetwork

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

Donors, members, sponsors

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Platinum sponsors

Microsoft · intel · SOCIETE GENERALE
Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE — MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION · HUAWEI

Gold sponsors

openinventionnetwork · Université de Paris · SORBONNE UNIVERSITÉ

Silver sponsors

CAST Software Intelligence for Digital Leaders · Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE · GitHub · Google · Università di Pisa · vmware

Bronze sponsors

DANS Data Archiving and Networked Services · FOSSID · ALMA MATER STUDIORUM A.D. 1088

Outline

1. Introductions
2. Software source code: a pillar of Open Science
3. Software Heritage
4. Under the hood: architecture and data structure
5. Demo time!
6. A revolutionary infrastructure
7. Building for the long term
8. Milestones
9. Breaking news
10. Conclusion

Some key dates

Summer 2015

The collection starts: first server, (very) early prototype

June 30th 2016

Public unveiling, with the first sponsors: Microsoft and DANS

April 3rd 2017

Unesco - Inria agreement on software access and preservation.

June 7th 2018

June 7, 2018

Opening the door of the Software Heritage archive

Opening the archive to the world

December 7th 2018

Starting the mirror network

February 26th 2019

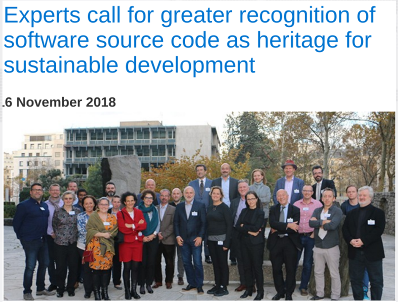
Part of our Heritage, Pillar of our Present, Enabler of our Future

Publication of the expert meeting

Paris Call on

Software Source Code

Policy highlight



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 …

United Nations Educational, Scientific and Cultural Organization

Memory of the World

Inria inventors for the digital world

PARIS CALL
SOFTWARE SOURCE CODE
AS HERITAGE FOR SUSTAINABLE DEVELOPMENT

Software Heritage

Their call is published on Feb 2019

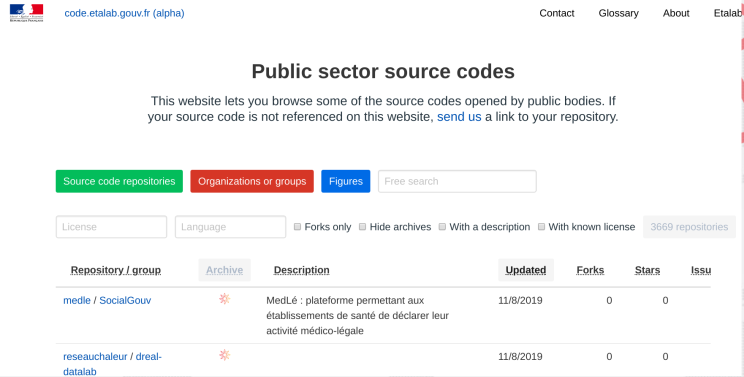
It's an important *policy tool*, already referenced and used … *yes, you can sign it!*

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

Outline

1. Introductions
2. Software source code: a pillar of Open Science
3. Software Heritage
4. Under the hood: architecture and data structure
5. Demo time!
6. A revolutionary infrastructure
7. Building for the long term
8. Milestones
9. **Breaking news**
10. Conclusion

Archiving *public* code



code.etalab.gouv.fr (alpha)

Contact Glossary About Etalab

Public sector source codes

This website lets you browse some of the source codes opened by public bodies. If your source code is not referenced on this website, send us a link to your repository.

□ Forks only □ Hide archives □ With a description □ With known license 3669 repositories

Repository / group	Archive	Description	Updated	Forks	Stars	Issu
medle / SocialGouv	✻	MedLé : plateforme permettant aux établissements de santé de déclarer leur activité médico-légale	11/8/2019	0	0	
reseauchaleur / dreal-datalab	✻		11/8/2019	0	0	

https://code.etalab.gouv.fr

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

The Software Heritage Acquisition Process (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

Adoption is coming ...

HAL software curated deposit workflow

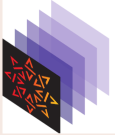
Curated Archiving of Research Software Artifacts
International Journal of Digical Curation, 2020

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

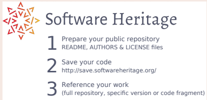
JTCAM (Theor. Comp. and Appl. Mech)

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

Policy

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)

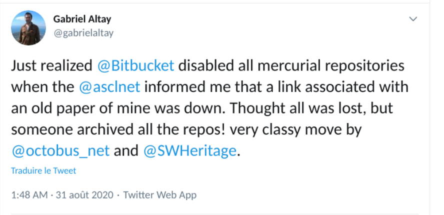
- online summary
- full ICMS 2020 paper

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)



Gabriel Altay
@gabrielaltay

Just realized @Bitbucket disabled all mercurial repositories when the @asclnet informed me that a link associated with an old paper of mine was down. Thought all was lost, but someone archived all the repos! very classy move by @octobus_net and @SWHeritage.

Traduire le Tweet

1:48 AM · 31 août 2020 · Twitter Web App

Bottomline

explicit deposit is important, ...

... and we must promote it...

... but will never be enough.

(think also of all software dependencies!)

You can help: 1/2

Towards Reproducible Open Science

archive research software in SWH

reference it using *intrinsic identifiers*

build on top of SWH, *do not try to rebuild SWH!*

reduce risk

avoid fragmentation!

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.

A *common* infrastructure

- mutualisation for sustainability
- open source, non for profit
- mirror network open to all
- let's prevent a useless diaspora

You can help: 2/2

Scientific challenges in building Software Heritage

graph queries, efficient storage, distributed archival, classification, search, …

Using Software Heritage for research

the *Software Heritage graph dataset* is now available!

- AWS: `https://registry.opendata.aws/software-heritage/`
- guidelines: `https://upsilon.cc/~zack/research/publications/msr-2019-swh.pdf`

Outline

1. Introductions
2. Software source code: a pillar of Open Science
3. Software Heritage
4. Under the hood: architecture and data structure
5. Demo time!
6. A revolutionary infrastructure
7. Building for the long term
8. Milestones
9. Breaking news
10. Conclusion

Join the revolution!

www.softwareheritage.org @swheritage

Library of Alexandria of code

- recover the past
- structure the future

A CERN for Software

- build better software
  - for industry
  - for society as a whole

Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchiroli
Building the Universal Archive of Source Code
Communications of the ACM, October 2018

Roberto Di Cosmo, Morane Gruenpeter, Stefano Zacchiroli
Identifiers for Digital Objects: the Case of Software Source Code Preservation
iPRES 2018: Intl. Conf. on Digital Preservation