# Software Heritage

## Building the infrastructure to track all software source code

Roberto Di Cosmo

roberto@dicosmo.org

January 11th, 2019
Software Intelligence Forum

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

# Outline

# 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

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

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

# Outline



# Software is eating the world…

## Business

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts

ESSAY

**Why Software Is Eating The World**

*By Marc Andreessen*

August 20, 2011

This week, Hewlett-Packard (where I am on the board) announced that it is exploring jettisoning its struggling PC business in favor of investing more heavily in software, where it sees better potential for growth. Meanwhile, Google plans to buy up the cellphone handset maker Motorola Mobility. Both moves surprised the tech world. But both moves are also in line with a trend I've observed, one that makes me optimistic about the future

Software companies

outperform or buy out

hardware companies

*Marc Andreesen, 2011*

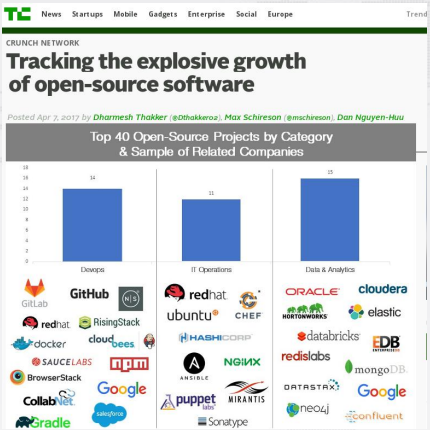
## Technology

**Software Defined Everything**

Hardware gets commoditised

Software becomes the new value!

# … Open Source is eating the Software World





## Open Source Software

can be openly (re)used, modified, (re)distributed, *with full access to its source code!*

# Reuse is the new rule

## Sonatype survey (2017)

80% to 90% of a new application is … just reuse!

## Where does reused software come from?


## Do *you* know where it comes from?

- the software you ship
- the software you use
- the software you acquire
- the software that
  - has that bug
  - has that vulnerability

## You should better know…

placeholder for off the record anecdotes

# Outline



# Software Heritage, in a nutshell

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

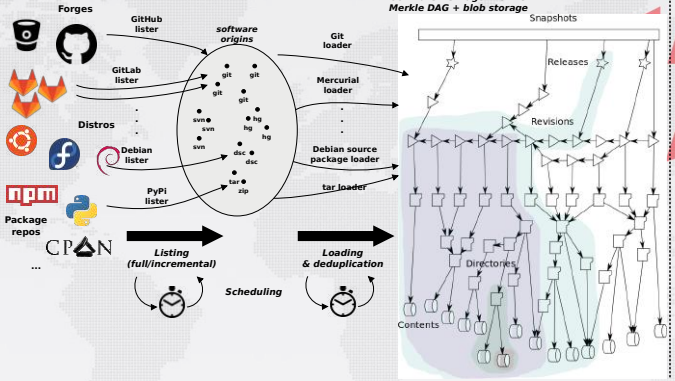
enable analysis of all the source code

# Coverage


GitHub debian GitLab Google code {P} python™ Package Index GITORIOUS GNU HAL archives-ouvertes.fr Inria inventeurs du monde numérique

- 200 TB (compressed) blobs, 6 TB database (as a graph: 10 B nodes + 100 B edges)
- The *richest* public source code archive, … and growing daily!

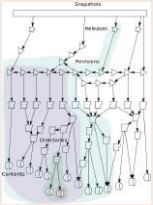
# Automation, and storage




- full development history permanently archived!

# A revolutionary infrastructure for Industry

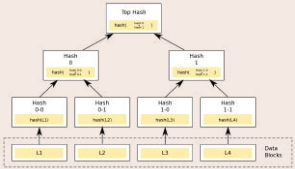
## The *graph* of Software Development




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)

# Outline

# Bird's eye view of compliance

## Customers

The "software industry" …

develops, maintains, uses, and/or integrates software components

## Needs

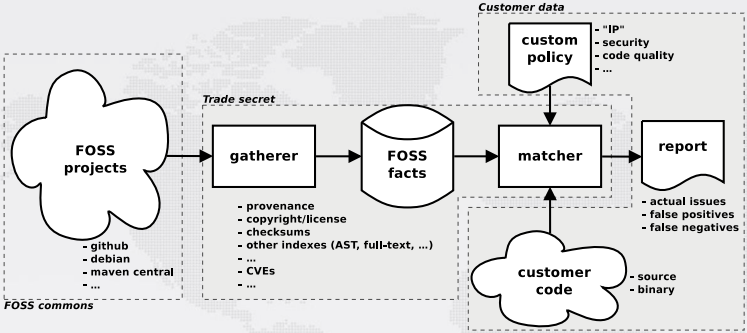
- track provenance of source code
- enforce licensing policies, avoid costly violations
- streamline supply chain management…

  acceptance process, subcontractors, etc.
- medium to long term software maintenance

## Tools

- code scanners
- bill of materials reporters

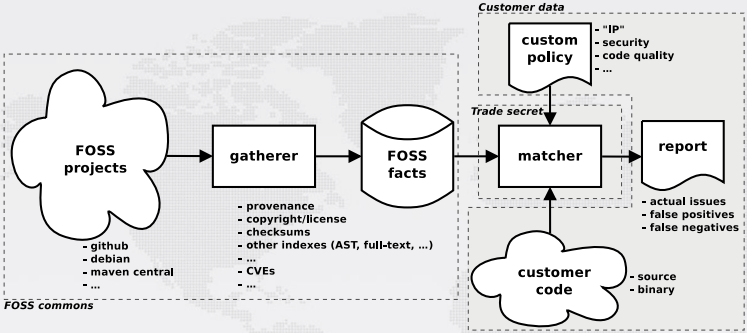
# The toolchain today




## Limitations

- non-free tools, secret data bases
  - poor (or no) reproducibility
  - low quality
  - no interoperability
  - sharing *off business core* knowledge is difficult

# The toolchain we want for tomorrow

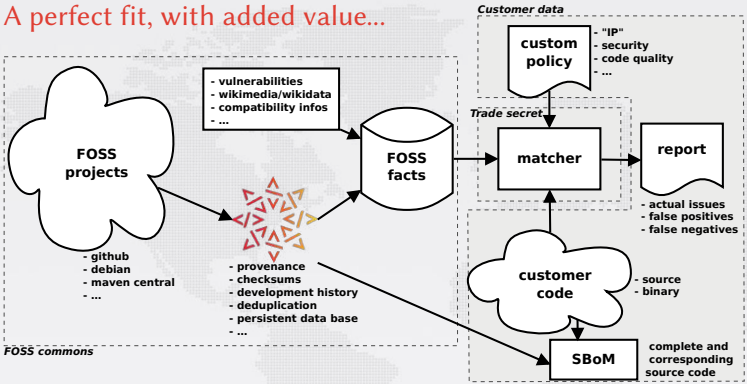



## Advantages

- shared effort
  - improved quality
  - reduced cost
- common, *transparent* knowledge base

# Software Heritage to the rescue

## A perfect fit, with added value...




## Get the complete corresponding source code

- even when upstream has gone away
- not just a legal obligation
- cornerstone for product maintenance and evolution

# Outline



# Raising Awareness

## April 3rd 2017, Unesco Inria agreement

Inria
INVENTEURS DU MONDE NUMÉRIQUE

UNESCO

## November 2018, Unesco Inria expert call

UNESCO
United Nations Educational, Scientific and Cultural Organization
"Building peace in the minds of
ABOUT US | THEMES | COUNTRIES | PARTNERSHIPS | JO

Home › All News › Experts call for greater recognition of software source code as heritage for sustainable development

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

**16 November 2018**

# Growing Support

## Sharing the vision


## Donors, members, sponsors


## Research collaboration


source code search engine

## Global network


- first independent mirror
- increased reliability

# Fresh news

## A new collaboration is in the works to:

- build a provenance index on Software Heritage
- let you answer questions like:
  - what is the first occurrence of this source file?
  - where else this source file was used?

In a nutshell, know your software

# Outline



# Come in, we're open!

Software Heritage

www.softwareheritage.org @swheritage

## Library of Alexandria of code

- recover the past
- structure the future

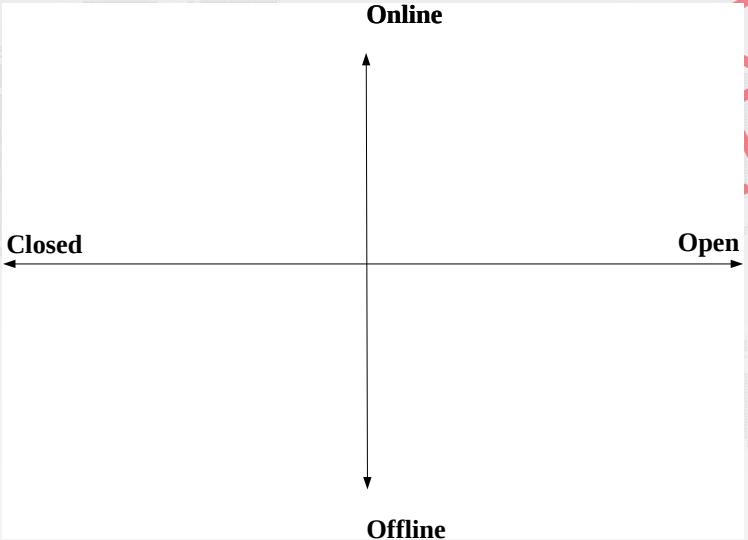
## A CERN for Software

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

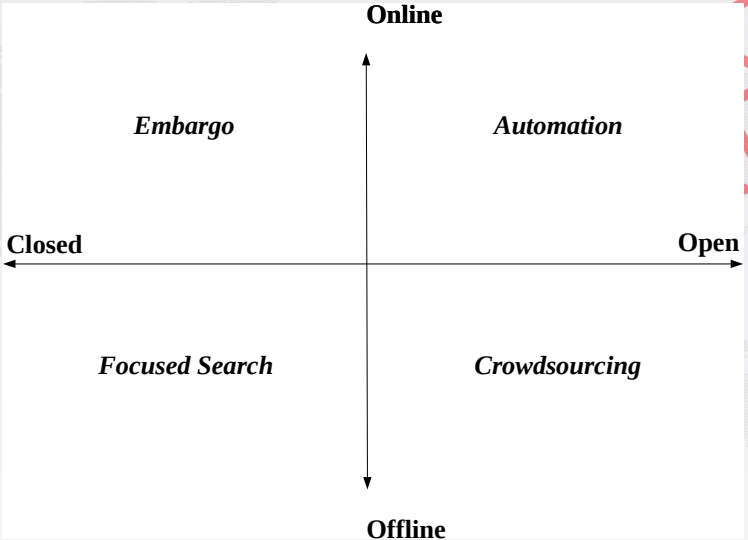
## Becoming a sponsor

https://sponsorship.softwarheritage.org

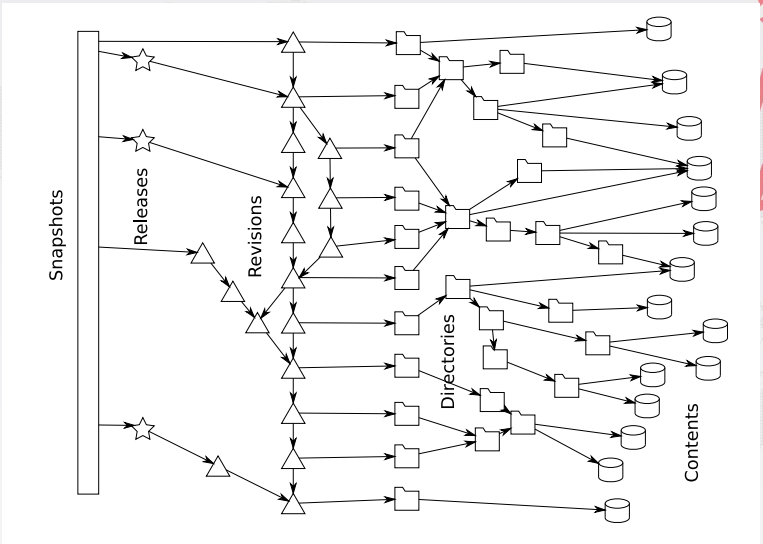
# All the source code



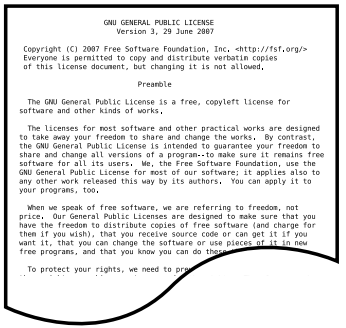
# All the source code: strategy



# 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.

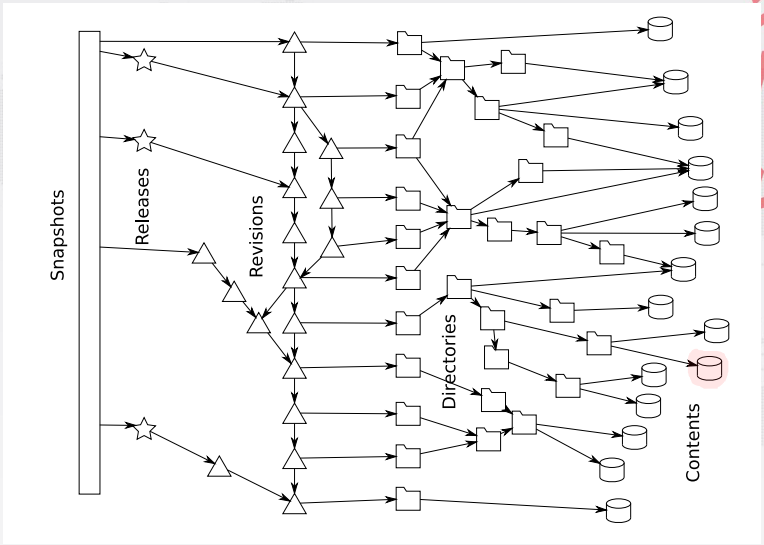
The licenses for most software and other practical works are designed
to take away your freedom to share and change the works. By contrast,
the GNU General Public License is intended to guarantee your freedom to
share and change all versions of a program--to make sure it remains free
software for all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

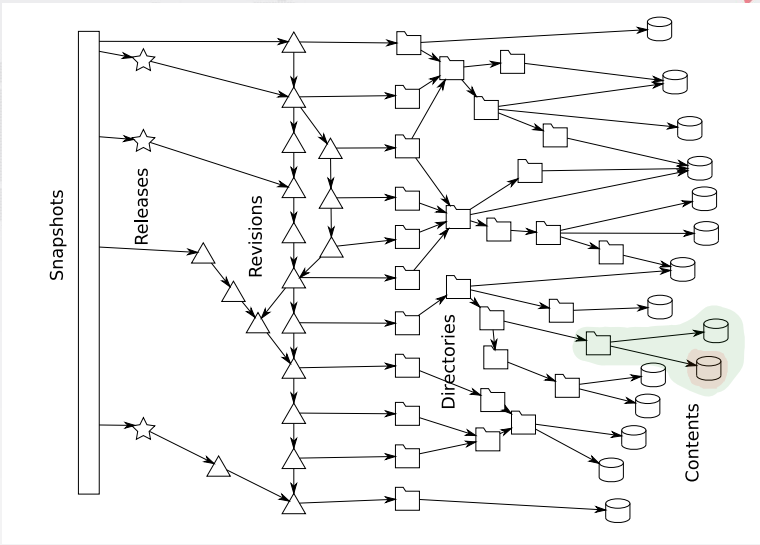



## Directories

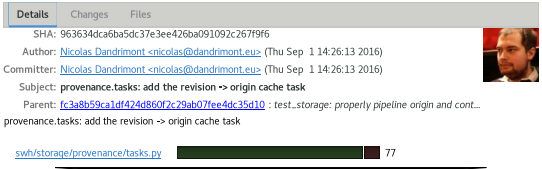
```
100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d0a34af6f52cf3cf6b0c2f7bd0648fbd255e77f AUTHORS
100644 blob 94a9ed024d3859793618152ea559a168bbcbb5e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e0867751dfb095c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5eaf2eb4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



# Revisions

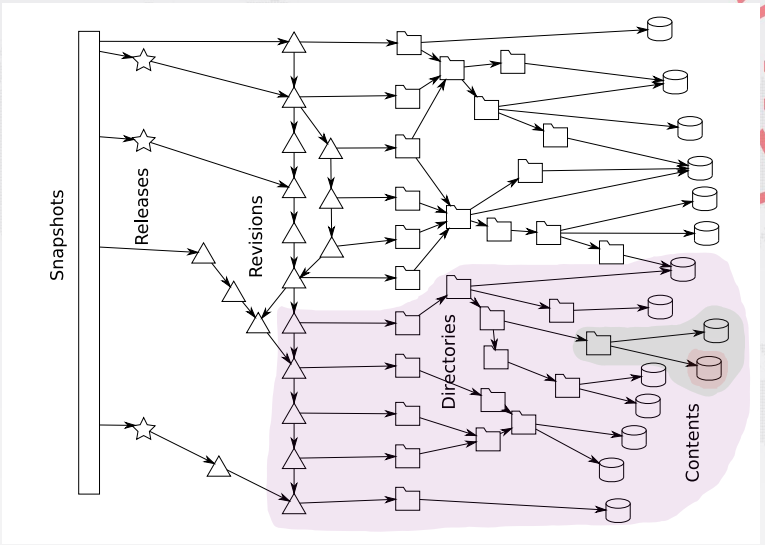



tree 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
parent fc3a8b59ca1df424d860f2c29ab07fee4dc35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

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

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6

# The archive in pictures



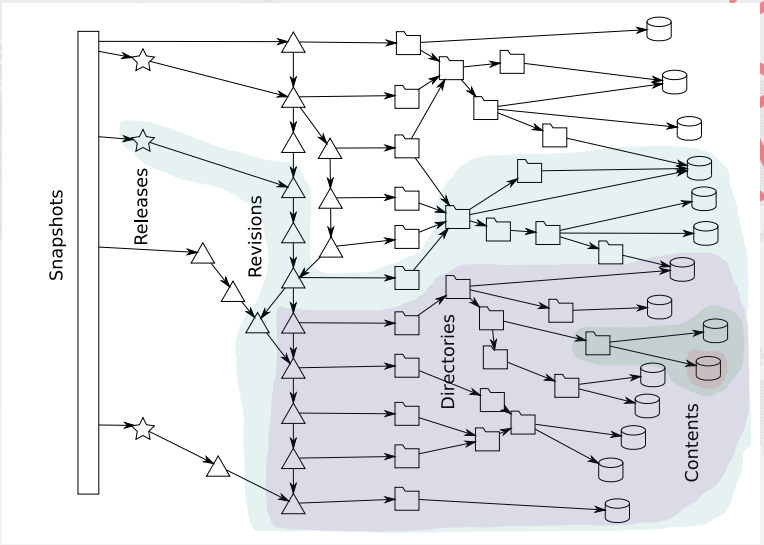
# Releases

```
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
ICBIlt2ujtXuCrDt93eKKPwvzZXq+hB0sMWy35Dr6jW7Z7K4Mu/PGglyIHPY55yo
IGEndWno7Vfh1Vm6t1n5qB7l5mXRaqA+becqddubTZ2xjj+jpIUqC8cyqN3hm/fL
qsj2mu8kyz3t8tG/H1/pV+I5OwBInPoS5TH0tujoJEVgPK/dHSP79QuHDHZFkCao
klj6kAWyU80Mxb+nKV/jeLbrR3+yWBFj3Qp5a1/V8oOTh6E1dALcNMpEaKCoKtMt
d/gMRax1ILg0EDfnsW67G6sDwKPKPHngfVLQ3nV3GaQQTnu1RpMz006N9/hAwzC
Gg/K1PdHT4hzOll46wYPZyje0U2VXGFu6vVU9vFQ4ZR/WJn+0zMzdcRdRJlSUOMn
RpTTfUsbXUeXHGOpkgXhSYTnvp1gdPc76USTsK0aGe84AZm1Ik0mGrwXCVfPqlYo
nhhibBSHBNMoqyF6yTSOpUbYK70tpYRRUGKWDeRK0wKSxkWKUZGtKzy6JYqJjo29
gulwgZQif5qWQCB0OontAL2+HvPFaVyckMejUhg62cP/+EHIvUk=
=kOxP
-----END PGP SIGNATURE-----
```

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb

# The archive in pictures



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 7eca197fc66d2024047e54b1ed9e8b4361a0fc2 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 3147c3d31ec46cf6492f801e908b1237ebdff2c7 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

# A bird's eye view