

Software Heritage

Building the Universal Software Archive for Open Science

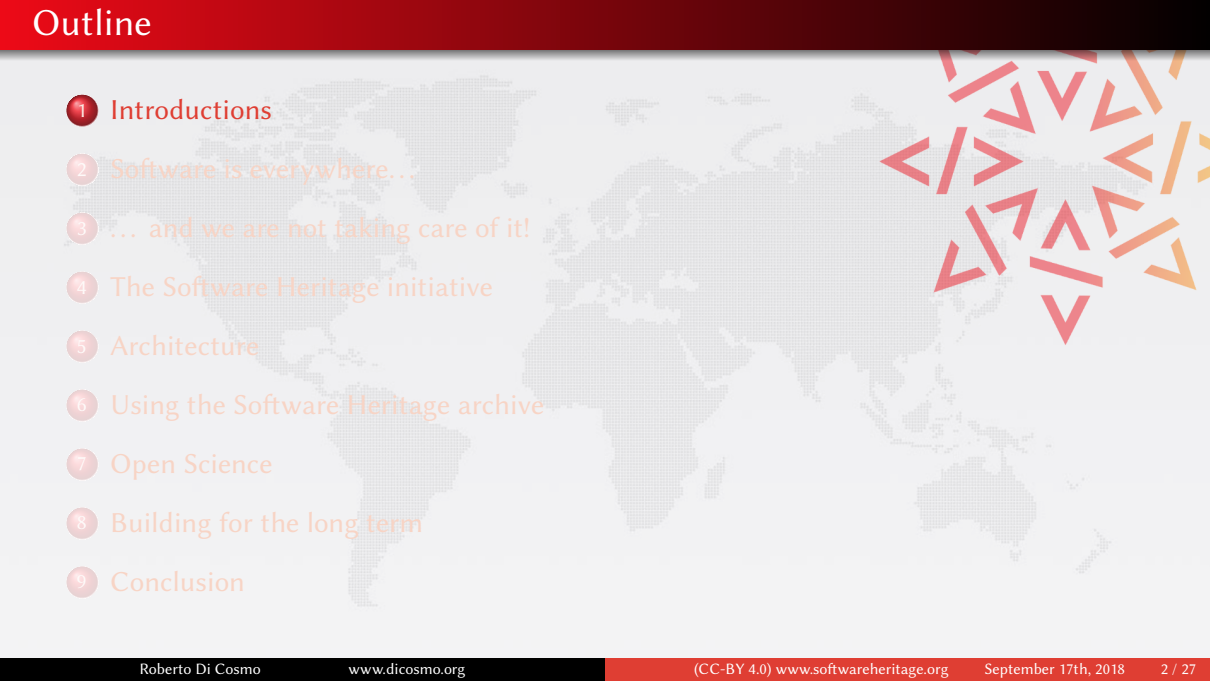
Roberto Di Cosmo

`roberto@dicosmo.org`

September 17th, 2018



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

- 
- 1 Introductions
 - 2 Software is everywhere...
 - 3 ... and we are not taking care of it!
 - 4 The Software Heritage initiative
 - 5 Architecture
 - 6 Using the Software Heritage archive
 - 7 Open Science
 - 8 Building for the long term
 - 9 Conclusion

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



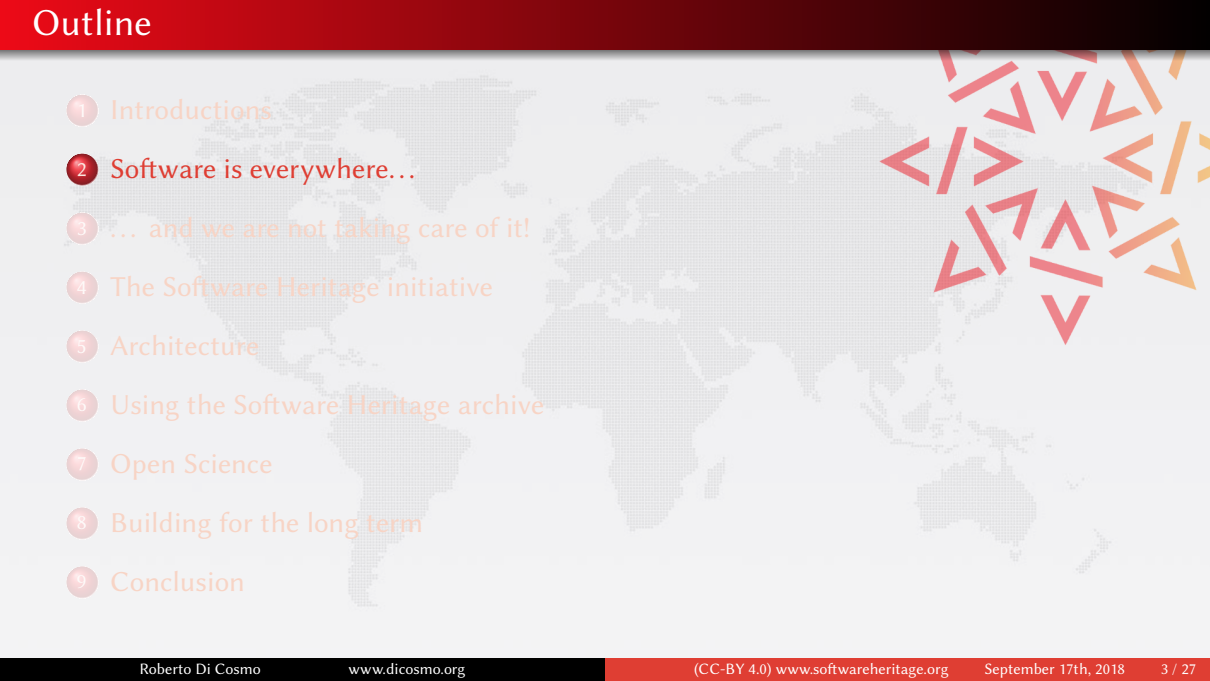
1999 *DemoLinux* – first live GNU/Linux distro

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

2008 *Mancoosi project* www.mancoosi.org

2010 *IRILL* www.irill.org

2015 *Software Heritage* at INRIA

- 
- 1 Introductions
 - 2 Software is everywhere...
 - 3 ... and we are not taking care of it!
 - 4 The Software Heritage initiative
 - 5 Architecture
 - 6 Using the Software Heritage archive
 - 7 Open Science
 - 8 Building for the long term
 - 9 Conclusion

Software is everywhere



Source code is *executable* and *human readable* knowledge

a growing part of our *Cultural Heritage*

Source code is *special*

Harold Abelson, Structure and Interpretation of Computer Programs

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

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;
}
```

Net. queue in Linux (excerpt)

```
/*
 * SFB uses two B[1][n] : L x N arrays of bins (L levels, N bins per level)
 * This implementation uses L = 8 and N = 16
 * This permits us to split one 32bit hash (provided per packet by rxhash or
 * external classifier) into 8 subhashes of 4 bits.
 */
#define SFB_BUCKET_SHIFT 4
#define SFB_NUMBUCKETS (1 << SFB_BUCKET_SHIFT) /* N bins per Level */
#define SFB_BUCKET_MASK (SFB_NUMBUCKETS - 1)
#define SFB_LEVELS (32 / SFB_BUCKET_SHIFT) /* L */

/* SFB algo uses a virtual queue, named "bin" */
struct sfb_bucket {
    u16      qlen; /* length of virtual queue */
    u16      p_mark; /* marking probability */
};
```

Len Shustek, Computer History Museum

“Source code provides a view into the mind of the designer.”

~ 50 years, a lightning fast growth

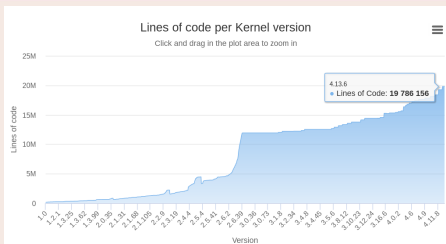
Apollo 11 Guidance Computer (~60.000 lines), 1969



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

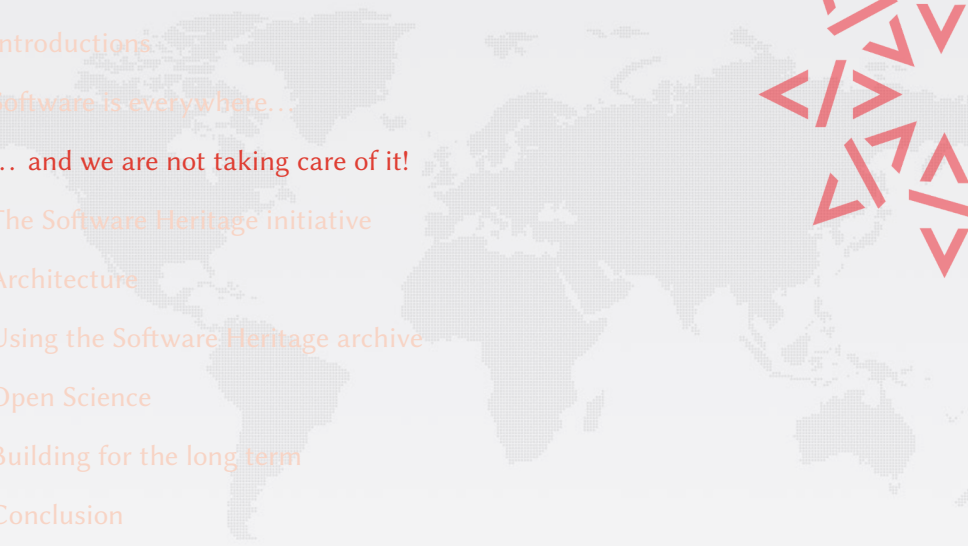
Margaret Hamilton

Linux Kernel



... now in your pockets!

are we taking care of all this?

- 
- 1 Introductions
 - 2 Software is everywhere...
 - 3 ... and we are not taking care of it!
 - 4 The Software Heritage initiative
 - 5 Architecture
 - 6 Using the Software Heritage archive
 - 7 Open Science
 - 8 Building for the long term
 - 9 Conclusion



Software is spread all around



A word cloud featuring various software-related terms and logos, including Sourceforge, Git, GitHub, Bitbucket, GoogleCode, Gitlab, CTan, CRAN, Debian, Inria, Maven, Gitorious, BerliOs, and Adullact. The words are arranged in a circular pattern, with some overlapping. The background is a light gray world map. To the right of the word cloud is a decorative graphic consisting of a series of red and orange arrows pointing outwards from a central point, forming a star-like shape.



A word cloud centered on a faint world map background. The words are arranged in a circular pattern, with 'damage' and 'disaster' being the largest. Other prominent words include 'malicious', 'obsolete', 'deletion', 'attack', 'format', 'corruption', 'dependencies', 'aging', 'tear', 'media', 'dangling', 'wear', 'encryption', 'reference', and 'storage'. The words are in various colors including purple, brown, blue, green, and red.

damage
disaster
malicious
obsolete
deletion
attack
format
corruption
dependencies
aging
tear
media
dangling
wear
encryption
reference
storage

Software lacks its own research infrastructure



Photo: ALMA(ESO/NAOJ/NRAO), R. Hills

Research software: a long way to go!

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

URL decay disrupts the *web of reference*

Web links *are not* permanent (even *permalinks*)

there is no general guarantee that a URL... which at one time points to a given object continues to do so

T. Berners-Lee et al. Uniform Resource Locators. RFC 1738.

404

URLs used in articles *decay*!

Analysis of *IEEE Computer* (Computer), and the *Communications of the ACM* (CACM): 1995-1999

- the *half-life* of a referenced URL is *approximately 4 years* from its publication date
D. Spinellis. The Decay and Failures of URL References.

Communications of the ACM, 46(1):71-77, January 2003.

Similar findings in Lawrence, S. et al. *Persistence of Web References in Scientific Research*, IEEE Computer, 34(2), pp. 26-31, 2001.

An example from Astronomy

Domain	links (broken)	.html	.txt	.dat	.gz	.tar	.fits	tilde
cxc.harvard.edu	802 (110)	336 (70)	0	0	4 (2)	5 (4)	1	0
heasarc.gsfc.nasa.gov	640 (33)	423 (27)	1	0	0	0	0	0
www.stsci.edu	498 (61)	205 (29)	3	0	0	0	0	15 (10)
asc.harvard.edu	471 (152)	212 (99)	0	0	0	0	0	1 (1)
ssc.spitzer.caltech.edu	427 (194)	125 (76)	3 (3)	0	0	0	0	0
cfa-www.harvard.edu	352 (68)	277 (52)	1	0	0	0	0	54 (17)
archive.stsci.edu	308 (58)	57 (9)	2	1 (0)	0	0	0	0
www.ipac.caltech.edu	285 (14)	209 (12)	0	0	0	0	0	0
www.atnf.csiro.au	211 (21)	12 (6)	0	0	0	0	0	7 (5)
space.mit.edu	193 (10)	58 (5)	1	0	0	0	0	2 (1)
www.astro.psu.edu	186 (4)	103 (1)	1	10	1	1	0	2
www.eso.org	186 (58)	54 (22)	1 (1)	0	0	0	0	4 (1)
irsa.ipac.caltech.edu	163 (5)	38	0	0	1	0	0	0
www.sdss.org	156 (2)	106 (1)	0	0	0	0	0	0
hea-www.harvard.edu	125 (37)	42 (17)	1	0	0	1	0	26 (16)
physics.nist.gov	125 (3)	63 (2)	0	0	0	0	0	0
www.noao.edu	120 (3)	50 (2)	0	0	0	0	0	0
xmm.vilspa.esa.es	118 (35)	23 (19)	0	0	8 (1)	0	0	1 (1)
www.astro.princeton.edu	115 (31)	43 (14)	0	0	0	0	0	53 (12)
ad.usno.navy.mil	110 (27)	98 (22)	3 (3)	0	0	0	0	1 (1)

This table lists total number of links and broken links (HTTP status codes 3xx, 4xx, and 5xx) to top domains (domains with over 100 links) found within articles published in the four main astronomy journals between 1997 and 2008. The table also shows, for each domain, the portion of links to common filename extensions, as well as links that contain the tilde character.
doi:10.1371/journal.pone.0104798.t001

How Do Astronomers Share Data?

Pepe, Goodman, Muench, Crosas, Erdmann

[dx.doi.org/10.1371/journal.pone.0104798](https://doi.org/10.1371/journal.pone.0104798)

PLOS August 28, 2014

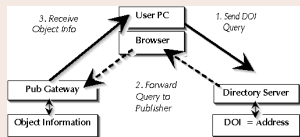
DOI limitations

Example: doi:10.1109/MSR.2015.10

- to find what 10.1109/MSR.2015.10 is, go to a *resolver* (e.g. doi.org)
- this returns <http://ieeexplore.ieee.org/document/7180064/>
- at this URL we find ...



Architecture of the DOI infrastructure



- DOI resolution *can change*
- content at URL *can change*
- no *intrinsic* way of noticing
- persistence based on *good will of multiple parties*

No catalog, no archive, no references: we are at a turning point

Looking at the past

- a lot of old software misplaced, lost, or behind barriers, but...
- most founding fathers are still here, and willing to share
- **urgent** to collect their knowledge

Only a few years left.

Looking at the future

- software development and use skyrockets: more programmers, and more code!
- **essential** to provide a **universal** platform for all the future software source code

Every year that goes by makes the problem worse.

it is **urgent** to take action!

- 
- 1 Introductions
 - 2 Software is everywhere...
 - 3 ... and we are not taking care of it!
 - 4 The Software Heritage initiative**
 - 5 Architecture
 - 6 Using the Software Heritage archive
 - 7 Open Science
 - 8 Building for the long term
 - 9 Conclusion



Software Heritage

Our mission

Collect, **preserve** and **share** the *source code of all the software* that is available

Past, present and future

Preserving the past, enhancing the present, preparing the future

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.

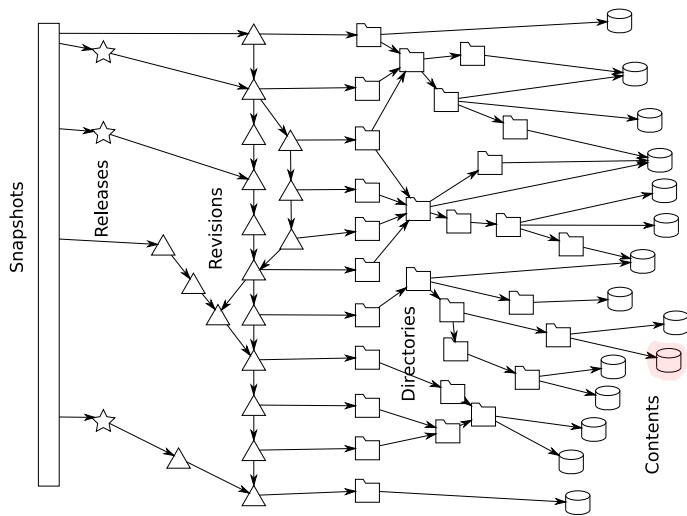
The licenses for most software and other practical works are designed
to take away your freedom to share and change the works. By contrast,
the GNU General Public License is intended to guarantee your freedom to
share and change all versions of a program--to make sure it remains free
software for all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

... To protect your rights, we need to prevent others from denying you
these rights by patenting or otherwise restricting the use of the software.

sha1: 8624bcdade55baeef...
sha256: 8ceb4b9ee5aded...
sha1_git: 94a9ed024d385...
length: 35147

The archive in pictures



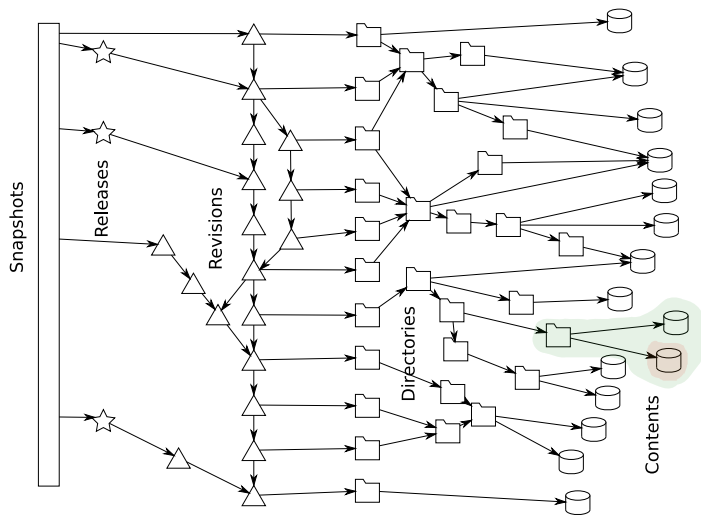


Directories

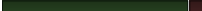
100644	blob	c5baade4c44766042186ef858c0fd63d587ebf09	.gitignore
100644	blob	2d0a34af6f52cf3cf6b0c2f7bd0648fbd255e77f	AUTHORS
100644	blob	94a9ed024d3859793618152ea559a168bbcbb5e2	LICENSE
100644	blob	d9b2665a435a43f8a79a84e0867751dfb095c7bb	MANIFEST.in
100644	blob	524175c2bad0b35b975f79284c2f5a6d5eaf2eb4	Makefile
100644	blob	5c7e3a5bbddb038682ba7793f440492ed9678bb3	Makefile.local
100644	blob	8617980629cd24e6080404f09aa749b085b3e07b	README.db_testing
100644	blob	76b29f94cf815e0869c414d38d78d7ce08ec514e	README.dev
040000	tree	e1e10ecf948af0b93adb0372afc89f12e92618a	bin
040000	tree	83e56d0beaf7793c77a45a345c80fcb8af503013	debian
040000	tree	a34c9c4ba213f0cedc67f9816348d27955577af5	docs
100644	blob	f2a6d32c6135aa7287bbd76167b01df2ae4f1539	requirements.txt
100755	blob	eee147c36caf1bbc2d820da8dc026cb5b68180bc	setup.py
040000	tree	224bb4c1f4c67fca1d160bfd2d06094e7e1abf3	sql
040000	tree	8631c9cd77bbe993168107ab5baf51f40c6300be	swl
040000	tree	8fb905b56ba8ed692f1209b2773b474c6c1d66c1	utils

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d

The archive in pictures



Revisions

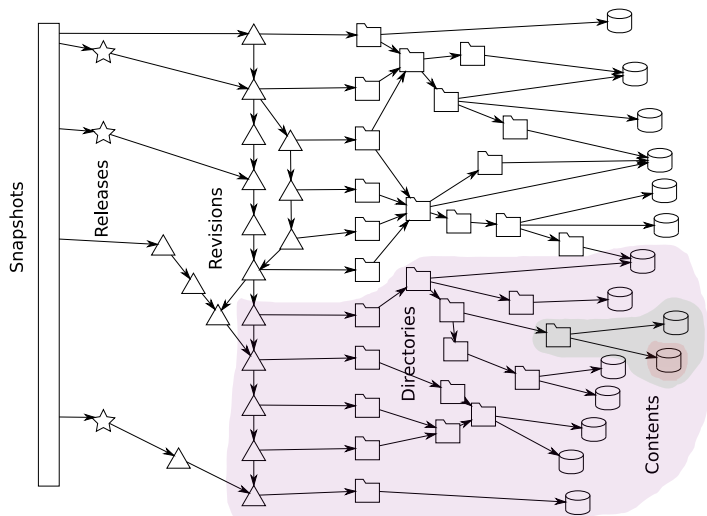
Details	Changes	Files
SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6		
Author: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)		
Committer: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)		
Subject: provenance.tasks: add the revision -> origin cache task		
Parent: 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

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 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

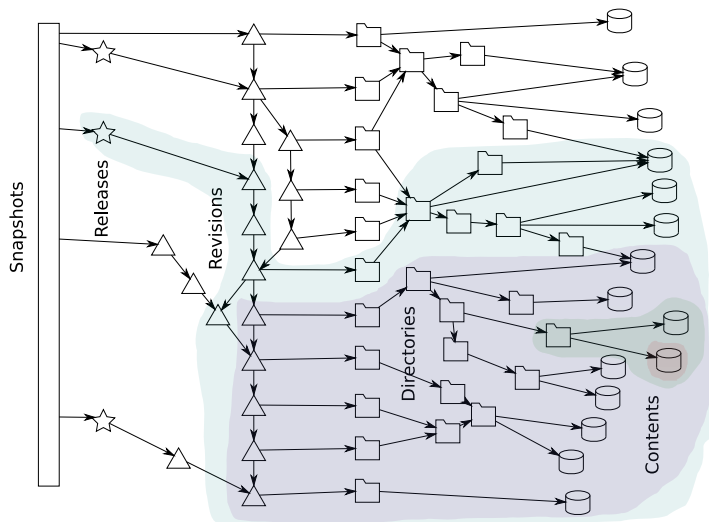
-----BEGIN PGP SIGNATURE-----

iQIzBAABCAAdBQJXvZTNFhxuaWNvbGFzQGRhbmRyaW1vbnQuZXUACgkQ7AWLMo2+
neqorw/aaq6SOb5DijzEa+kWN3rXgV5+1K1vEVh1wNKAwx8eKJ7aX2kEiLDtt7uf
ahpZ6pz3q8nqs6aC1+YrxBfcih3L2YtrdZeXWwqr8xWNMaEoYDb8qaphwh8AD5t2
ICBlit2ujtXuCrDt93eKKPwvZxg+hB0sMWy35Dr6jW7Z7K4Mu/PGglylHPY55yo
IGEndWno7VfH1Vm6t1n5qB7l5mXRaqa+becqddubTZ2xjj+jpIUqC8cyqN3hm/fL
qsj2mu8kyz3t8tG/H1/pV+I5OwBlNpO5STH0tuoJEvGPK/dHSP79QuHDHDFkCao
klj6kAWyU80Mxb+nKV/jelBrR3+yWBFj3Qp5a1/V8oOTh6E1dALcNMpEaKCoKtMt
d/gMRax1l1/g0EDfnsW67G6sDwKPKPHgfvLQ3nV3GaQQTru1RpMz006H9/tAwzC
Gg/K1PdHT4hzOIl46wYPZyje0U2VXGFu6vVU9vFQ4ZR/Wjn+0zMzdcRdrJlSUOMN
RpTfUlsbXUeXHGOpkgXhSYTrnp1gdPc76U5TsK0aGeB4AZm1lk0mGrwXCVFpqlYo
nhhibB5HBNMoqyF6yTSOpUbyK70tpYRRUGKWDेरK0wKSxkWKUzGtKzy6jYqIjo29
gulwvZQif5qWQCB0OontAL2+HvPFaVycKMejUhg62cP/+EHlvUk=
=kOxP

-----END PGP SIGNATURE-----

id: [85083a5cc14a441c89dea73f5bdf67c3f9c6afdb](#)

The archive in pictures



git show-refs

Snapshots

```
commit 08ffeb25770109525eb3ce21691466c53a1d9158 refs/heads/atime
commit ba5443a24e3f9fe323a46c292cecf4cbe61c67eb refs/heads/directory-listing-arrays
commit d69e0dbf892383ff6589b27fbelc05d27238d9c5 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 215ea50daba11e082e0b72e76eb4b6073a87908 refs/tags/v0.0.21
tag 3fb168c2072a5d6252124257a1e5dfc0f5ffa1df refs/tags/v0.0.22
tag 8cdbee8da4d73fc5d262789e460a16ac3c72aba4 refs/tags/v0.0.23
...
```

id: b464cad1b66fff266a37b46ea6e7a04b545e904b

Cultural Heritage



Industry



Research



Education



Software Heritage

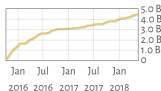
Source files

4,536,067,027



Commits

1,024,675,748



Projects

83,801,775



Technology

- transparency and FOSS
- replicas all the way down

Content

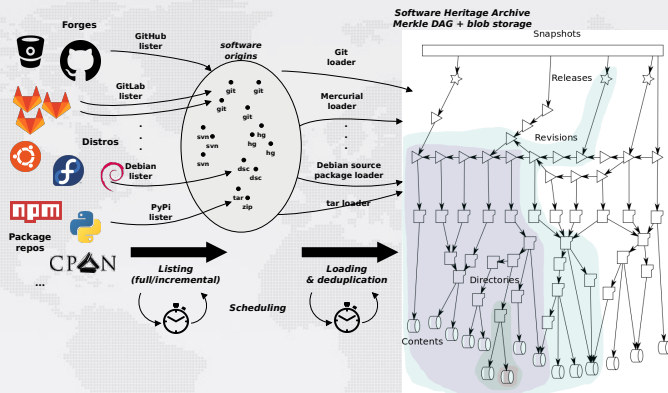
- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- mirror network

- 
- 1 Introductions
 - 2 Software is everywhere...
 - 3 ... and we are not taking care of it!
 - 4 The Software Heritage initiative
 - 5 Architecture**
 - 6 Using the Software Heritage archive
 - 7 Open Science
 - 8 Building for the long term
 - 9 Conclusion

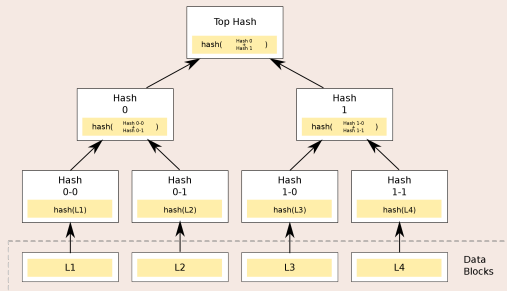
Automation, and storage



- full development history permanently archived
- origins: GitHub (auto), Debian (auto), **Gitlab.com**, Gitorious, Google Code, GNU
- ~ **200Tb** raw contents, ~ **10Tb** graph (**10Bn nodes**, **100Bn edges**)

Much more than an archive!

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



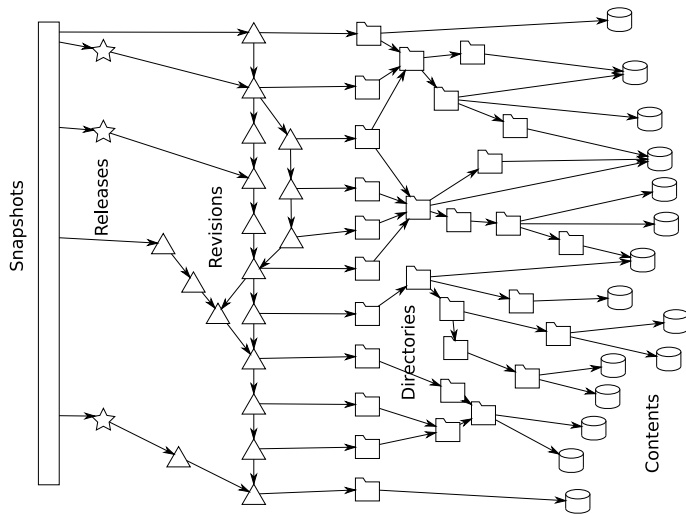
Combination of

- 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



- 
- 1 Introductions
 - 2 Software is everywhere...
 - 3 ... and we are not taking care of it!
 - 4 The Software Heritage initiative
 - 5 Architecture
 - 6 Using the Software Heritage archive**
 - 7 Open Science
 - 8 Building for the long term
 - 9 Conclusion

Reference archive for all software

A "wayback machine" for software source code ...

with **intrinsic identifiers**!

- <http://archive.softwareheritage.org/browse>
- <http://bit.ly/swhpids> for persistent identifiers

Demo time: let's highlight some features...

Origin search

The screenshot shows the 'Search' page of Software Heritage. A search bar at the top contains the text 'search softwareheritage.org to browse'. Below the search bar, a table lists search results. The table has columns for 'Origin type', 'Origin identifier', and 'Exclusion'. The results are filtered to show 'Software' origins. The first result is 'Browsersync@browsersync.org:latest', which is excluded. The second result is 'Browsersync@browsersync.org:latest', which is not excluded. The third result is 'Browsersync@browsersync.org:latest', which is not excluded. The fourth result is 'Browsersync@browsersync.org:latest', which is not excluded. The fifth result is 'Browsersync@browsersync.org:latest', which is not excluded. The sixth result is 'Browsersync@browsersync.org:latest', which is not excluded. The seventh result is 'Browsersync@browsersync.org:latest', which is not excluded. The eighth result is 'Browsersync@browsersync.org:latest', which is not excluded. The ninth result is 'Browsersync@browsersync.org:latest', which is not excluded. The tenth result is 'Browsersync@browsersync.org:latest', which is not excluded.

Directory browsing

The screenshot shows the 'Directory' page of Software Heritage. The URL bar shows 'https://archive.softwareheritage.org/browse/https://github.com/lynnal/lynnal'. The page title is 'SWH object: Directory'. Below the title, there is a table listing files and directories. The table has columns for 'File', 'Mode', 'Size', and 'SHA-256'. The first row is 'lib', which is a directory. The second row is 'lib', which is a directory. The third row is 'lib', which is a directory. The fourth row is 'lib', which is a directory. The fifth row is 'lib', which is a directory. The sixth row is 'lib', which is a directory. The seventh row is 'lib', which is a directory. The eighth row is 'lib', which is a directory. The ninth row is 'lib', which is a directory. The tenth row is 'lib', which is a directory.

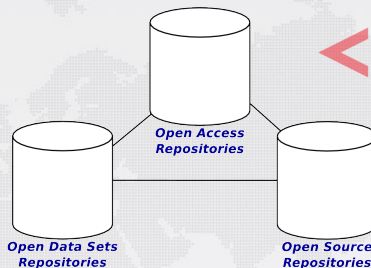
Revisions as diffs

The screenshot shows the 'Revisions' page of Software Heritage. The URL bar shows 'https://archive.softwareheritage.org/browse/https://github.com/lynnal/lynnal'. The page title is 'SWH object: Revisions'. Below the title, there is a table listing revisions. The table has columns for 'Revision', 'Size', and 'SHA-256'. The first row is '1', which is a revision. The second row is '2', which is a revision. The third row is '3', which is a revision. The fourth row is '4', which is a revision. The fifth row is '5', which is a revision. The sixth row is '6', which is a revision. The seventh row is '7', which is a revision. The eighth row is '8', which is a revision. The ninth row is '9', which is a revision. The tenth row is '10', which is a revision.

- 1 Introductions
- 2 Software is everywhere...
- 3 ... and we are not taking care of it!
- 4 The Software Heritage initiative
- 5 Architecture
- 6 Using the Software Heritage archive
- 7 Open Science**
- 8 Building for the long term
- 9 Conclusion



Supporting more accessible and reproducible science



A global library referencing all software used in all research fields

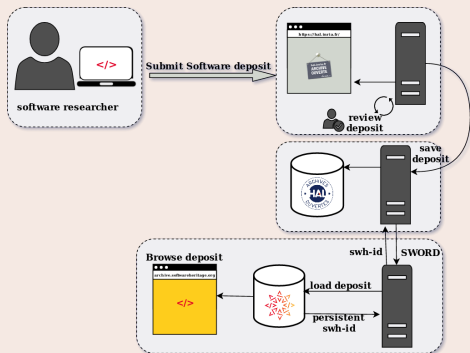
- completes the infrastructure for **Open Access** in science
- provides **intrinsic persistent identifiers** for scientific **reproducibility**
- enables large scale, verifiable **software studies**

Paper points to lost gitorious.org repo saved in Software Heritage

- https://www.openaire.eu/search/publication?articleId=dedup_wf_001::cd996f0b6236b90659f84f99feb62bcc
- <https://gitorious.org/parmap>
- <https://archive.softwareheritage.org/browse/search/?url=%22gitorious.org/parmap%22>

Deposit software in HAL

<http://hal.inria.fr/hal-01738741>



Generic mechanism:

- SWORD based
- review process
- versioning

How to do it:

- **today:** deposit .zip or .tar.gz file (*guide*)
- **tomorrow:**
 - provide *SWH id* and metadata
 - include *metadata file* for automatic metadata extraction
 - ...

September 2018: **open to all** on <https://hal.archives-ouvertes.fr/>

The way to go to archive and reference scientific software

All features of Software Heritage *for free*

- **intrinsic IDs** (integrity, not dependent on resolvers!)
 - specification: <http://bit.ly/swHPIDs>
 - **iPres2018** paper: <http://bit.ly/swHPIDpaper>
- browse, download (now)
- metadata, licenses, provenance (plagiarism detection), classification (wip), ...

Coverage and uniformity

- **one** archive for **all** domains (industry included)
- reference *any* software, not just the deposited ones
- **git-compatible** identifiers greatly simplify workflows


Sustainability

... doors are open!

one infrastructure

independent non profit foundation

worldwide mirrors

- 
- 1 Introductions
 - 2 Software is everywhere...
 - 3 ... and we are not taking care of it!
 - 4 The Software Heritage initiative
 - 5 Architecture
 - 6 Using the Software Heritage archive
 - 7 Open Science
 - 8 Building for the long term**
 - 9 Conclusion

Growing Support

Landmark Inria Unesco agreement, April 3rd, 2017



Sharing the vision



Contributing to the mission



The next steps

The Software Heritage Foundation

- independent
- long term mission
- multistakeholder

The community

- academia: Open Access, research
- industry: better software
- cultural heritage: **all** the software history

The mirror network

- resilience
- biodiversity

“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.”

Thomas Jefferson

You can help!

Many scientific and technological challenges

object storage, machine learning, classification, efficient graph queries, mirror protocols,
...

Contribute

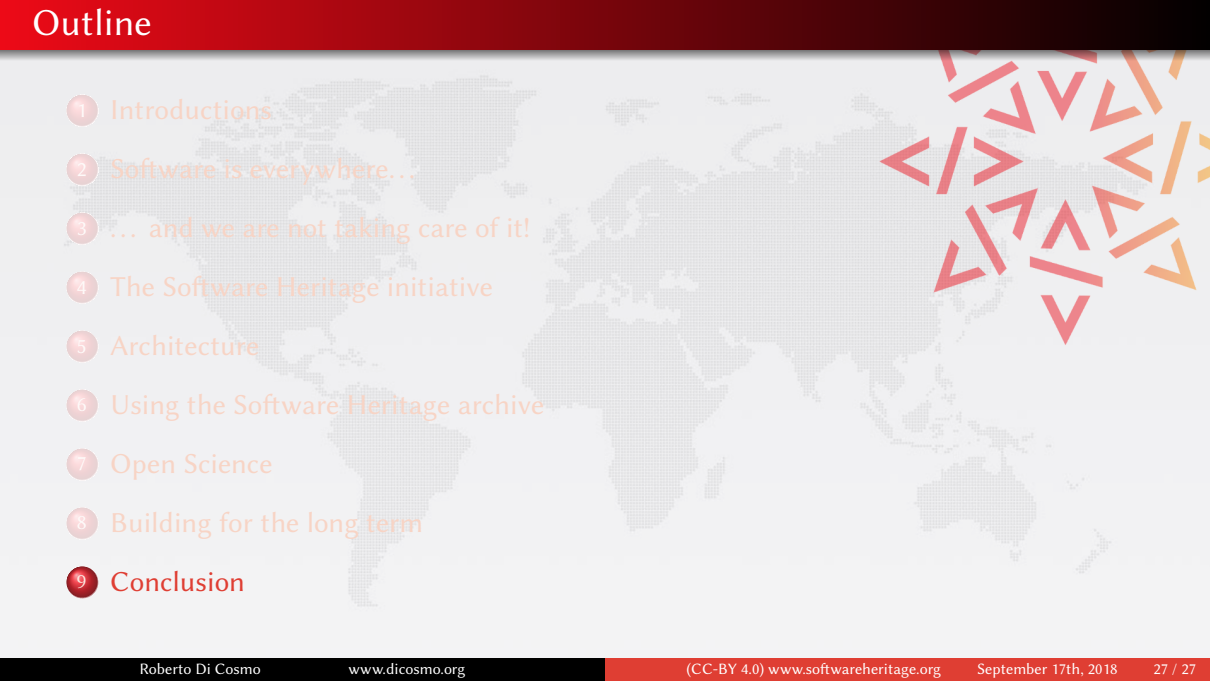
- `forge.softwareheritage.org`

Funding

- become a partner/sponsor/mirror :
`sponsorship.softwareheritage.org`
- give *your own contribution* :
`www.softwareheritage.org/donate`

Spread the word!

- *use* the archive and help others do
- tell everybody about Software Heritage

- 
- 1 Introductions
 - 2 Software is everywhere...
 - 3 ... and we are not taking care of it!
 - 4 The Software Heritage initiative
 - 5 Architecture
 - 6 Using the Software Heritage archive
 - 7 Open Science
 - 8 Building for the long term
 - 9 Conclusion**

Come in, we're open!



Software Heritage

www.softwareheritage.org

@swheritage

Library of Alexandria of code

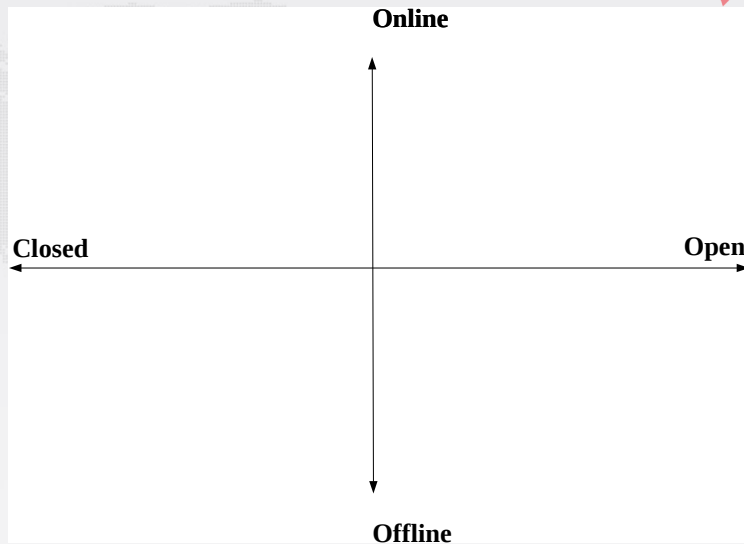


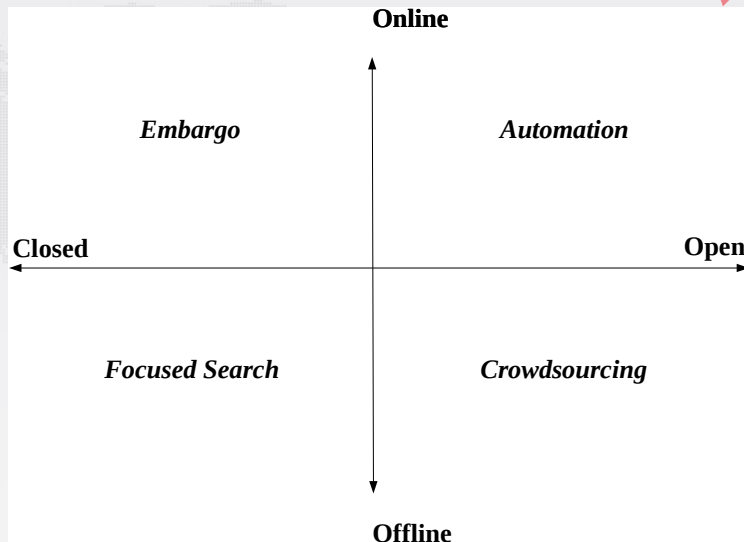
- recover the past
- structure the future

A CERN for Software

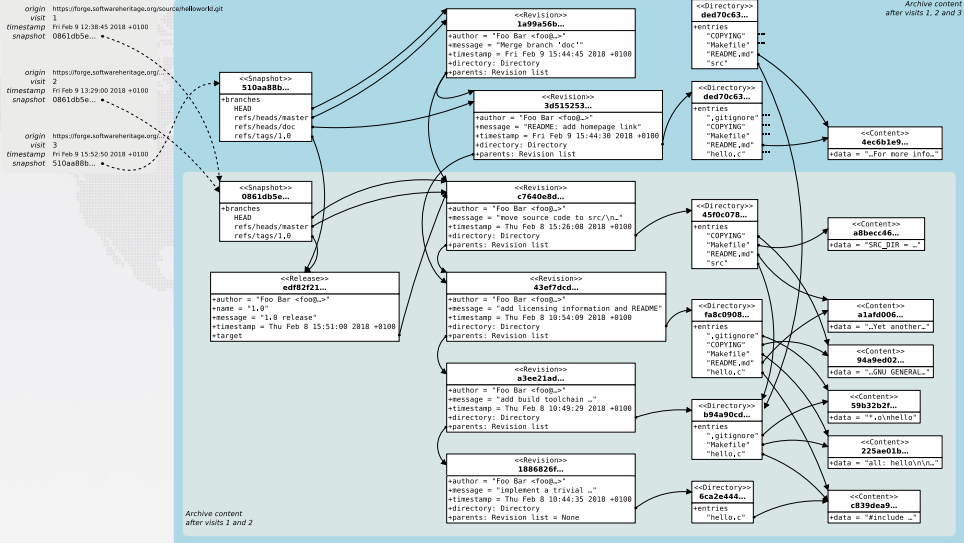


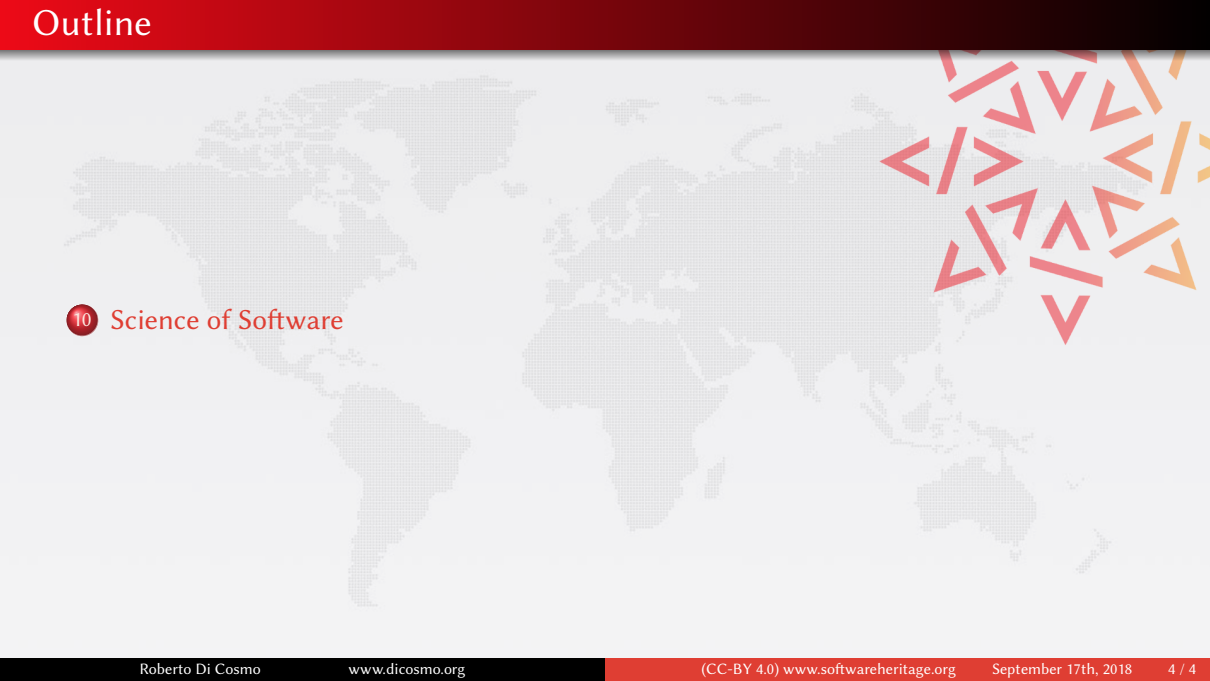
- build better software
 - for industry
 - for society as a whole





A bird's eye view





10 Science of Software



Large scale *repeatable* software studies...

- vulnerability detection
- dependency analysis
- pattern elicitation
- automatic classification ...

... need a uniform representation

Software Heritage has **one data model** for all forges/VCS...

... yes, we do **data normalization** of software evolution!

Breaking news: *soon* an **Amazon public data set!**