

Identifiers for Digital Objects

The Case of Software Source Code Preservation

Roberto Di Cosmo, Morane Gruenpeter, Stefano Zacchiroli

`roberto@dicosmo.org`

September 25th, 2018



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Source Code: *executable* and *human readable* knowledge



"The source code for a work means the preferred form of the work for making modifications to it."

GPL Licence

Hello World

Program (excerpt of binary)

```
4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
4004f9: 90
4004fa: 5d
4004fb: c3
```

Program (source code)

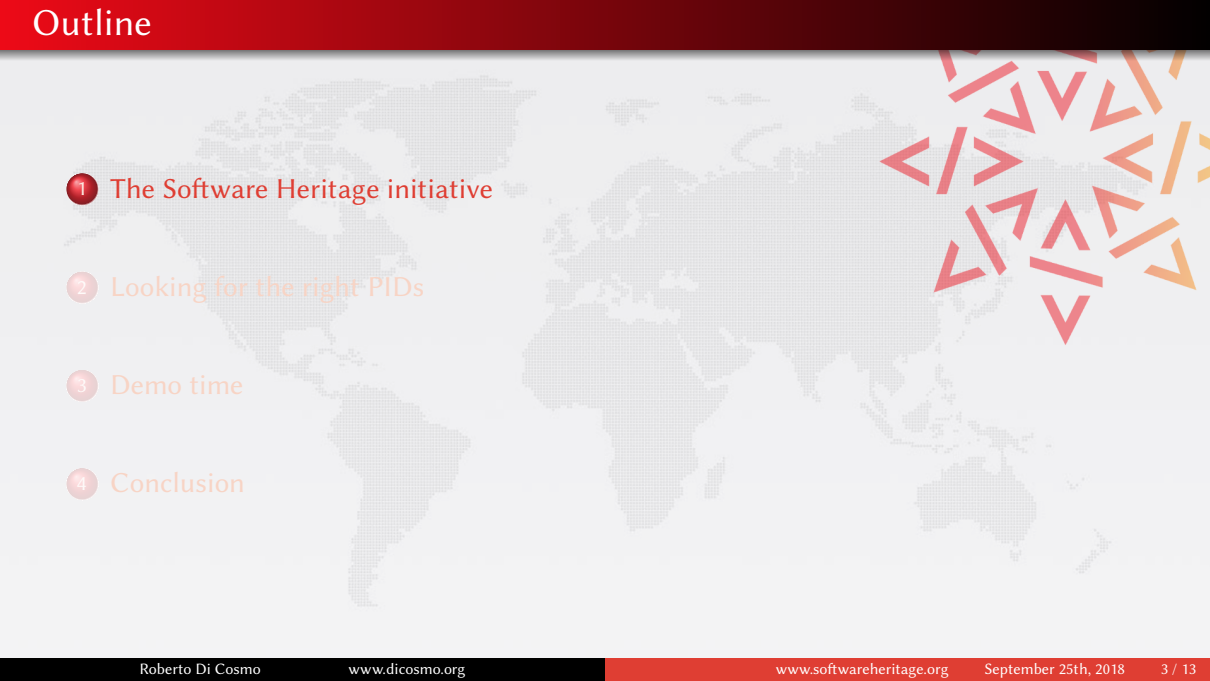
```
/* Hello World program */

#include<stdio.h>

void main()
{
    printf("Hello World");
}
```

Len Shustek, CHM

"Source code provides a view into the mind of the designer."

- 
- 1 The Software Heritage initiative
 - 2 Looking for the right PIDs
 - 3 Demo time
 - 4 Conclusion



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

Cultural Heritage



Industry



Research



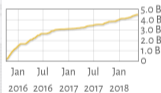
Education



Software Heritage

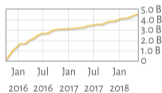
Source files

4,536,067,027



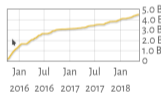
Commits

1,024,675,748



Projects

83,801,775



Technology


- transparency and FOSS
- replicas all the way down

Content (billions!)

- **intrinsic identifiers**
- facts and provenance

Organization

- non-profit
- multi-stakeholder

- 
- 1 The Software Heritage initiative
 - 2 Looking for the right PIDs
 - 3 Demo time
 - 4 Conclusion

A *system of identifiers* is

- a set of labels (the identifiers)
- mechanisms to perform :

<i>Generation (minting)</i>	create a new label
<i>Assignment</i>	associate label to object
<i>Retrieval</i>	get object from a label

- optionally, mechanisms to perform:

<i>Verification</i>	check label and object
<i>Reverse Lookup</i>	get label from an object
<i>Description</i>	get metadata of an object

Mechanisms offered in some systems of identifiers

Mech. / System	Handle	DOI	Ark	PURL
Generation	Yes	Yes	Yes	Yes
Assignment	Yes	Yes	Yes	Yes
Retrieval	Yes	Yes	Yes	Yes
Verification	N.A.	N.A.	N.A.	N.A.
Reverse Lookup	N.A.	N.A.	N.A.	N.A.
Description	Yes	Yes	Yes	N.A.

Our challenges in the PID landscape

Typical properties of systems of identifiers

uniqueness, non ambiguity, persistence, abstraction (opacity)

Key needed properties from our use cases

gratis identifiers are free (billions of objects)

integrity the associated object cannot be changed (sw dev, *reproducibility*)

no middle man no central authority is needed (sw dev, *reproducibility*)

we could not find systems with both **integrity** and **no middle man** !

An important distinction: DIOs vs. IDOs

The term “Digital Object Identifier” is construed as “digital identifier of an object,” rather than “identifier of a digital object”
Norman Paskin. 2010

DIO (Digital Identifier of an Object)

digital identifiers for (potentially) **non digital objects**

- epistemic complexity (manifestations, versions, locations, etc.)
- need an authority to ensure persistence and uniqueness

IDO (Identifier of a Digital Object)

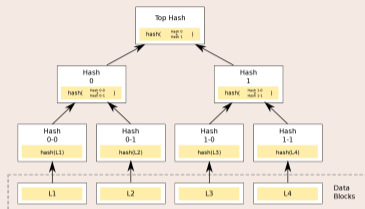
digital identifiers (only) for **digital objects**

- can provide both **integrity** and **no middle man**
- broadly used in modern software development (git, etc.)

for the core Software Heritage archive, **IDOs are enough**

IDO in Software Development: the origins

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



Combination of

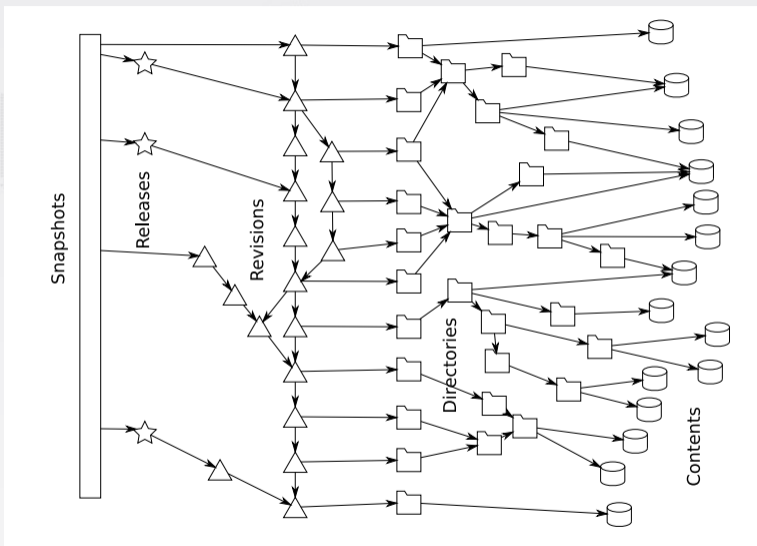
- tree
- hash function

Classical cryptographic construction

fast, parallel signature of large data structures, built-in deduplication

- satisfies all three criteria: **gratis, integrity, no middle man!**
- widely used in industry (e.g., Git, nix, blockchains, IPFS, ...)

IDO in Software Heritage: a worked example



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.

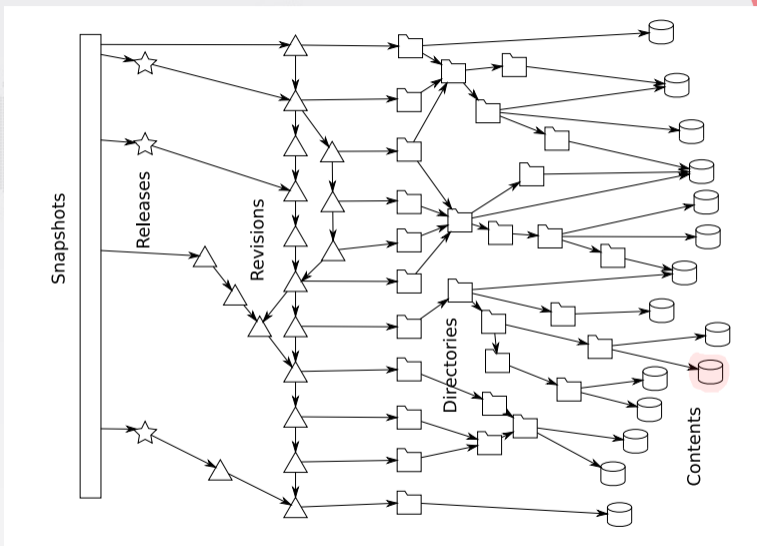
The licenses for most software and other practical works are designed
to take away your freedom to share and change the works. By contrast,
the GNU General Public License is intended to guarantee your freedom to
share and change all versions of a program--to make sure it remains free
software for all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

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

IDO in Software Heritage: a worked example



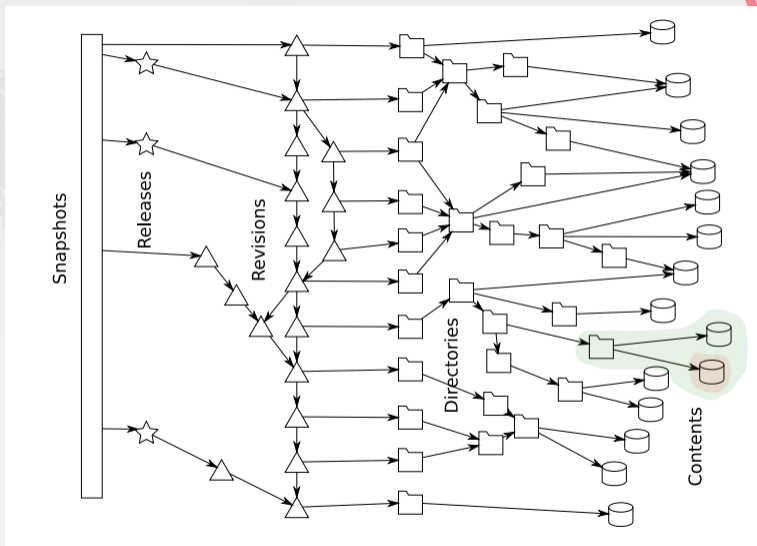


Directories


```
100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d0a34af6f52cf3cf6b0c2f7bd0648fbd255e77f AUTHORS
100644 blob 94a9ed024d3859793618152ea559a168bbcbb5e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e0867751dfb095c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5eaf2eb4 Makefile
100644 blob 5c7e3a5bbddb038682ba7793f440492ed9678bb3 Makefile.local
100644 blob 8617980629cd24e6080404f09aa749b085b3e07b README.db_testing
100644 blob 76b29f94cf815e0869c414d38d78d7ce08ec514e README.dev
040000 tree e1e10ecef948af0b93adb0372afc89f12e92618a bin
040000 tree 83e56d0beaf7793c77a45a345c80fcb8af503013 debian
040000 tree a34c9c4ba213f0cedc67f9816348d27955577af5 docs
100644 blob f2a6d32c6135aa7287bbd76167b01df2ae4f1539 requirements.txt
100755 blob eee147c36caf1bbc2d820da8dc026cb5b68180bc setup.py
040000 tree 224bb4c1f4c67fca1d160bfd2d06094e7e1abf3 sql
040000 tree 8631c9cd77bbe993168107ab5baf51f40c6300be swh
040000 tree 8fb905b56ba8ed692f1209b2773b474c6c1d66c1 utils
```

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d

IDO in Software Heritage: a worked example



Revisions

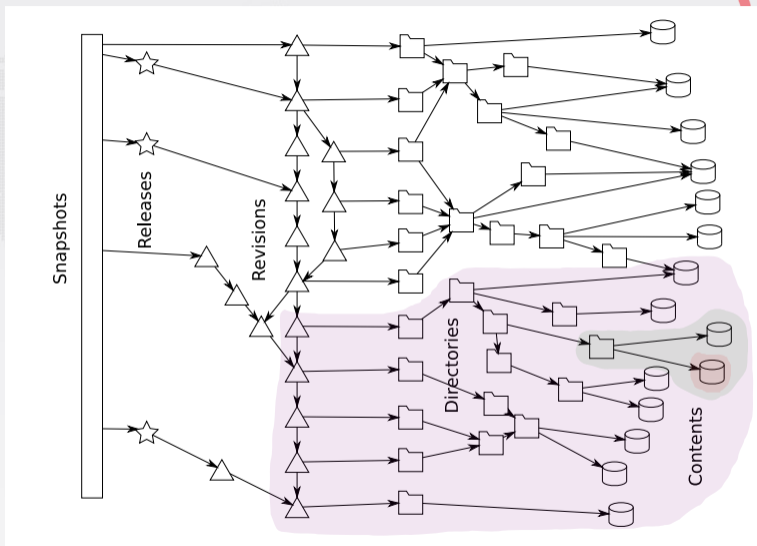
Details	Changes	Files
SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6		
Author: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)		
Committer: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)		
Subject: provenance.tasks: add the revision -> origin cache task		
Parent: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

IDO in Software Heritage: a worked example



Releases

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

Release swh.storage v0.0.51

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

commit c0c9f16b1e134f593e7567570a1761b156e6eb1d

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

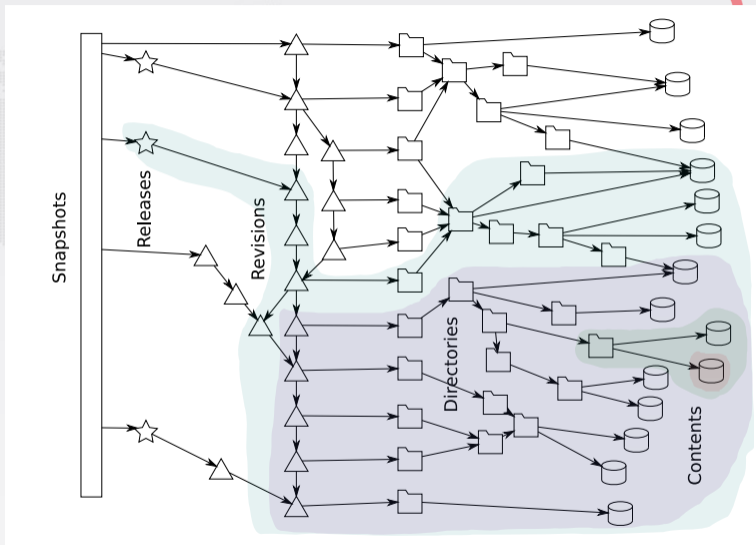
Release swb.storage v0.0.51

- Add new metadata column to origin_visit
- Update swb-add-directory script for updated API
---BEGIN PGP SIGNATURE---

```
iQIzBAABCAAdBQJXvZTNFhXuaWNvbGFzQGRhbmRyaW1vbnQuZXUACgkQ7AWLMo2+
neqorw/aaq65Ob5DijzEa+kWN3rXgV5+1K1vEVh1wNKAw8eKJ7aX2kEiLdt7uf
ahpZ6pz3q8nqs6aC1+YrxBfcih3L2YtrdZeWXXWqr8xWNMaEoYDb8qaphwh8AD5t2
ICBlit2ujtXuCrDt93eKKPwvzZxg+h80sMWy35Dr6jW7Z7K4Mu/PgGlyLHPY55yo
IGEndWno7VfH1Vm6t1n5qB7l5mXRaqA+becqddubTZ2xij+jpIUqC8cyqN3hm/fL
qsj2mu8kyz3t8tG/H1/pV+I5OwBlNpO5STH0tujojEVgPK/dHSP79QuHDHZFkCao
kij6kAWyU80Mxb+nKVjleLbrR3+yWBFj3Qp5a1/V8o0Th6E1dALcNMpEaKCoKtMt
d/gMRax1l1/g0EDfnsW67G6sDwKPKPhngfVLQ3nV3GaQQTnu1RpMz006H9/tAwzC
Gg/K1PdHT4hz0iI46wYPZyje0U2VXGFu6vVU9vFQ4ZR/Wjn+0zMzdcRdrJlSUOMn
RpTTfU5bXUeXHGOpkgXhSYTnvp1gdPc76U5TsK0aGe84AZm1Ik0mGrwXCvFPqYo
nhhbB5HBNMoqyF6yTSOpUbyK70tpYRRUGKwDeRk0wKSxkWKUZGtKzy6jYqJJo29
gulwgZQif5qWQC80OontAL2+HvFfaVyckMejUhg62cP/+EHlvUk=
=kOxP
---END PGP SIGNATURE---
```

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb

IDO in Software Heritage: a worked example



Snapshots

git show-refs

```
commit 08ffeb25770109525eb3ce21691466c53a1d9158 refs/heads/atime
commit ba5443a24e3f9fe323a46c292cec4fcbe61c67eb refs/heads/directory-listing-arrays
commit d69e0dbf892383ff6589b27fbelc05d27238d9c5 refs/heads/foo
commit c77f9eeea0eb22f8946908f5a8019f67de468e08 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 586fba4e580b4f5fab05f599367643cbbc1a9c7f refs/tags/v0.0.15
tag 8cd8b885f4098bf36317742bd289f660e5be51c refs/tags/v0.0.16
tag a542444ee3f0fbcd35efb202fee035c809abc7d6 refs/tags/v0.0.17
tag 228a2f1650dd12222e556559462e1e06fc4993d9 refs/tags/v0.0.18
tag 606979a4ca05d497fc0d24aad0dce82636ef47c refs/tags/v0.0.19
tag 32bf5a59fc2a323baa6d5f15a6ad5382ec275a67 refs/tags/v0.0.2
tag 3147c3d31ec46cf6492f881e908b1237ebdff2c7 refs/tags/v0.0.20
tag 215ea50daball1e082e0b72e76eb4b6073a87908 refs/tags/v0.0.21
tag 3fb168c2072a5d6252124257a1e5dfc0f5ffa1df refs/tags/v0.0.22
tag 8cdbee8da4d73fc5d262789e460a16ac3c72aba4 refs/tags/v0.0.23
...
```

id: b464cad1b66fff266a37b46ea6e7a04b545e904b

The Software Heritage IDO schema (see <http://bit.ly/swhpids>)

swh:1:**cnt**:94a9ed024d3859793618152ea559a168bbcbb5e2 full text of the GPL3 license

swh:1:**dir**:d198bc9d7a6bcf6db04f476d29314f157507d505 Darktable source code

swh:1:**rev**:309cf2674ee7a0749978cf8265ab91a60aea0f7d

a **revision** in the development history of Darktable

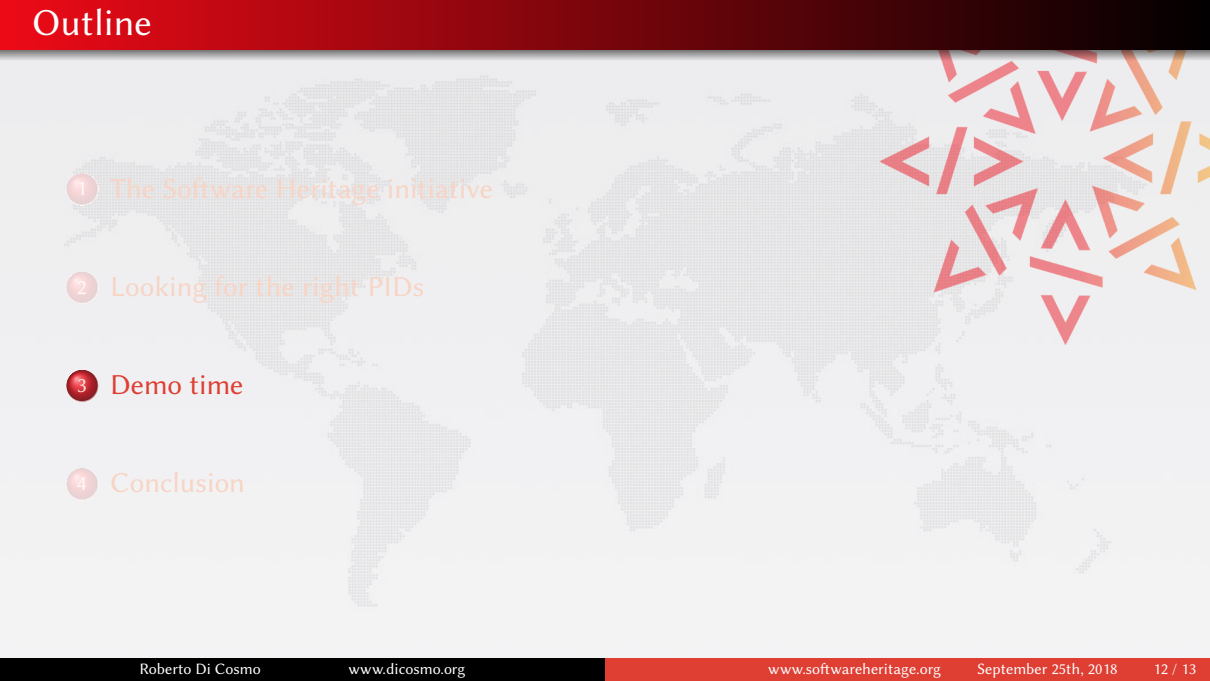
swh:1:**rel**:22ece559cc7cc2364edc5e5593d63ae8bd229f9f

release 2.3.0 of Darktable, dated 24 December 2016

swh:1:**snp**:c7c108084bc0bf3d81436bf980b46e98bd338453

a **snapshot** of the entire Darktable repository (4 May 2017, GitHub)

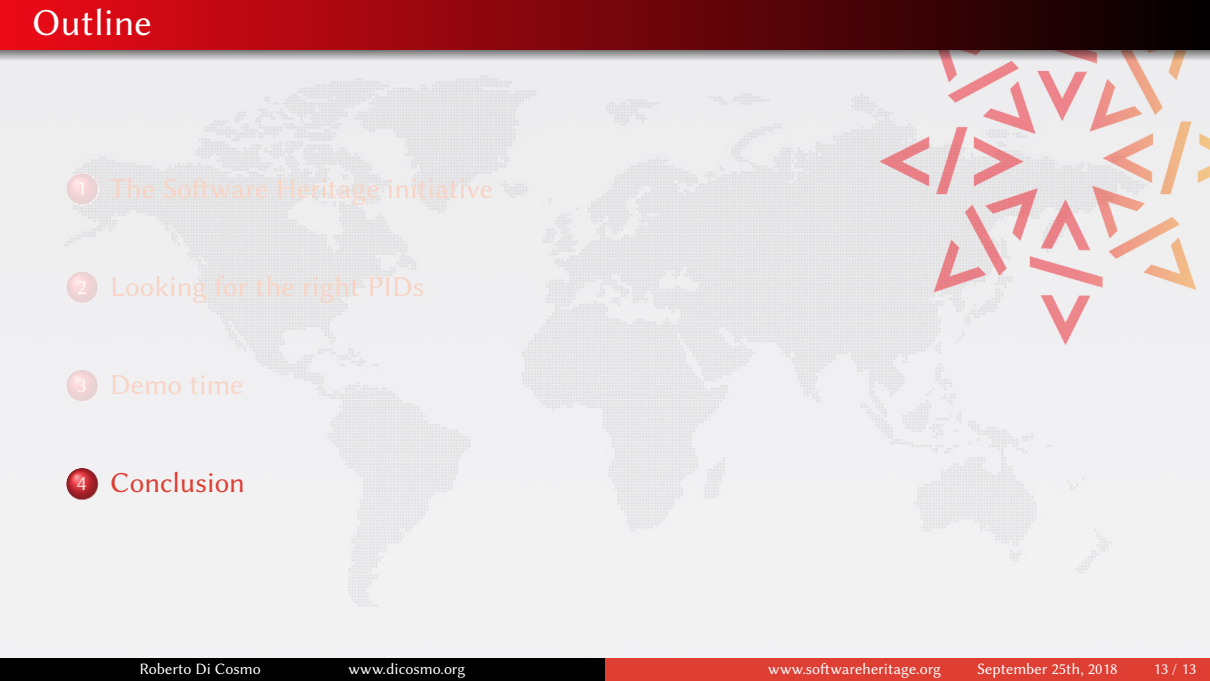
Current resolvers: archive.softwareheritage.org and n2t.org

- 
- 1 The Software Heritage initiative
 - 2 Looking for the right PIDs
 - 3 Demo time
 - 4 Conclusion

A "wayback machine" for software source code

Identifiers in action

- <http://archive.softwareheritage.org/browse>

- 
- 1 The Software Heritage initiative
 - 2 Looking for the right PIDs
 - 3 Demo time
 - 4 Conclusion

- there are many systems of identifiers
- DIOs and IDOs cater to different needs
- IDOs enable **integrity** and **no middle man** properties **together**
- Software Heritage is using IDOs for billions of objects, **today**
- we believe IDOs are appropriate for most **digital born** content that has a **canonical** representation

Come in, we're open!

www.softwareheritage.org – learn more

www.softwareheritage.org/support/sponsors/ – sponsoring info

www.softwareheritage.org/support/partners – partners

forge.softwareheritage.org – our own code

Questions?