Identifiers for Digital Objects
The Case of Software Source Code Preservation

Roberto Di Cosmo, Morane Gruenpeter, Stefano Zacchiroli

roberto@dicosmo.org

September 25th, 2018
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.”

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.”
Outline

1. The Software Heritage initiative
2. Looking for the right PIDs
3. Demo time
4. Conclusion
Software Heritage in a nutshell

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
A principled infrastructure


Software Heritage

Technology

- transparency and FOSS
- replicas all the way down

Content (billions!)

- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- multi-stakeholder
Outline

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:
  
<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generation</td>
<td>create a new label</td>
</tr>
<tr>
<td>Assignment</td>
<td>associate label to object</td>
</tr>
<tr>
<td>Retrieval</td>
<td>get object from a label</td>
</tr>
</tbody>
</table>

- optionally, mechanisms to perform:
  
<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verification</td>
<td>check label and object</td>
</tr>
<tr>
<td>Reverse Lookup</td>
<td>get label from an object</td>
</tr>
<tr>
<td>Description</td>
<td>get metadata of an object</td>
</tr>
</tbody>
</table>
Mechanisms offered in some systems of identifiers

<table>
<thead>
<tr>
<th>Mech. / System</th>
<th>Handle</th>
<th>DOI</th>
<th>Ark</th>
<th>PURL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Assignment</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Retrieval</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Verification</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>Reverse Lookup</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>Description</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>N.A.</td>
</tr>
</tbody>
</table>
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
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, …)
IDOs in Software Heritage: a worked example
Contents

GNU GENERAL PUBLIC LICENSE
Version 3, 29 June 2007
Copyright (C) 2007 Free Software Foundation, Inc. <http://gsf.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 make restrictions that prevent others from doing what will break these conditions. This License, the General Public License, applies to any software for which the copyright holder chooses it. Such software has been chosen to Public Domain ones described in Section 7. We have designed this License to be suitable for use a wide variety of software.

sha1: 8624bcdae55baeef...
sha256: 8ce4eb9ee5ad6ed...
sha1_git: 94a9ed024d385...
length: 35147
Directories

```
id: 515f00d44e92c65322aaa9bf3fa097c00dd9c7d
```
Revisions

<table>
<thead>
<tr>
<th>Details</th>
<th>Changes</th>
<th>Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHA: 953634dca6ba5dc37e3ee426ba091092c267f9f6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author: Nicolas Dandrimont <a href="mailto:nicolas@dandrimont.eu">nicolas@dandrimont.eu</a> (Thu Sep 1 14:26:13 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Committer: Nicolas Dandrimont <a href="mailto:nicolas@dandrimont.eu">nicolas@dandrimont.eu</a> (Thu Sep 1 14:26:13 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject: provenance.tasks: add the revision -&gt; origin cache task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent: fc3a8b59ca1df424d860f2c29ab07fee4d4c35d10</td>
<td></td>
<td>77</td>
</tr>
<tr>
<td>provenance.tasks: add the revision -&gt; origin cache task</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
tree 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
taxparent fc3a8b59ca1df424d860f2c29ab07fee4d4c35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
commiter Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

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

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6
Releases

```object
c0c9f16b1e134f593e7567570a1761b156e6eb1d
type comit
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-----

iQizBAABAACAd0qJXvZT NFhxuaWNvbfGFqOGRhbmNyawW1vbnQzuXUAcgkO74WLMo2 +
neq8wv/aq6S0o5Dj2e+a+KWN3drXpVS+1k1vEvn1wrAwx8ek7jx2kEeIDt1juf
ahp26zugq8pq6aCa+1xYr8CicJL2Ytd2ZeXwWMAe07YD8bqaphwhBAJ212
ICB92u1jxUCd93eKFpwcZxg+hB6oMYW35D6w7W7K4MwJOPgyHFPY5yv+
IGEndWno7V7H1Vm6t1n5pB7t5mXRaaoAg+beqcwd5ub7Zzxy+jqoUq8CpynQ3hvnF
q5jZmknzu2xI8Bgl/H1pV/=sOwPm0SoS5HotujuOEvP/kDhJSF7QqUj2H15Fcap
k1j4A4wVU8b63w6+bwJyULbr3+yW8bJyQy3g6a1/VWx6761eALCmMo3eAc0OZx
jQrMa3j2jo6EFdnoW67G656dXwKtKPhgjVQJO63V3GqAQQTNv1/Rpm0z46E94LauraC
gkX13pHT7h4oOf3wF7yoeU2VXQFu6Uv9fV3ZKWWi=2MzdcRdrrJ5UO4Mn
RyT1nubXeA7KQqpx9hs3TYnvp3dpC2yUS5rGGo6242JZmXk6mGwNCVPhqtb
nNh138DBHmMqopy4F6Y735pJyB7K0zyPc6U6KwGowkW5XZC675uyKa029
guWQZTif5WQCB0o3m9L2-RvFav5y0cMejU7hgb2cF7+E1tvUkE==
=KdXp
----END PGP SIGNATURE-----

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb
```
Snaphots

commit 08fbeb25770109525e03cc2169346c53a3d9158 refs/heads/atlime
commit ba4433a4e3d9fe023a36e292ce4f0c861c57eb refs/heads/directory-listing-arrays
commit d6e96db0f82383ff5689027be1c85d2732389d6c5 refs/heads/tno
commit cf77f9ea8e8b22f0840f0f08450f07e4556e888 refs/heads/master
commit 7e19219f8cf9d2b2404c4e8a541a5edc8e1bb4a4c61aacf2 refs/heads/tmp-directory-add
commit 642428f53780856b64b5d27b33ee4f12a252e82e refs/heads/tmp/generic-releases
tag 29f0943b13787f684965677799f644907c75f755 refs/tags/v0.1
>tag 72a2199a3b384e399599dd6b678fb/f8bbee732ae2c2d refs/tags/v0.1.10
>tag 3590e0ca0b6b8b8f8be03b376d5f3b59bfa4f3a5c refs/tags/v0.1.11
>tag 33378427a493ba5e6a7777b8d856674f1b556 refs/tags/v0.1.12
>tag 057465275b327c7590311c26a936c3f34035d refs/tags/v0.1.13
>tag 5a6352f8e89ab8454d368424d6679d2a13e2f3b6b refs/tags/v0.1.14
>tag 586f9ab4e5b0b47f5ab85f9367643cbca19c77 refs/tags/v0.1.15
>tag 8c088b85f109ebf383317777428f696d0ee5b51c refs/tags/v0.1.16
>tag a04a444e36f9be35e7f2f04fe93c899abc73d6 refs/tags/v0.1.17
>tag 278a2f16650d12222e556559402e160e6f4c993d9 refs/tags/v0.1.18
>tag 695979a4ca854497f8c024aadd9d0e6255eef47c refs/tags/v0.1.19
>tag 32b5a55f6cf2a233ba6654f103a653d32ce375207e7 refs/tags/v0.2
>tag 3147cd33c4c64f922ff881e9008b1237ediff2c7 refs/tags/v0.2.20
>tag 2135e5a5daba11e882e6072e7e64b6e873a87908 refs/tags/v0.2.21
>tag 3b16c927a0d8d222122310b0d4f4855f8d47d refs/tags/v0.2.22
>tag 8cbee8da4d73fc8247289e4e09a16ac3c72aba4 refs/tags/v0.2.23

id: b464cad1b66f266a37b46ea6e7a04b545e904b

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>swh:1:cnt:94a9ed024d3859793618152ea559a168bbcb5e2</td>
<td>full text of the GPL3 license</td>
</tr>
<tr>
<td>swh:1:dir:d198bc9d7a6bcf6db04f476d29314f157507d505</td>
<td>Darktable source code</td>
</tr>
<tr>
<td>swh:1:rev:309cf2674ee7a0749978cf8265ab91a60aea0f7d</td>
<td>a revision in the development history of Darktable</td>
</tr>
<tr>
<td>swh:1:rel:22ece559cc7cc2364edc5e5593d63ae8bd229f9f</td>
<td>release 2.3.0 of Darktable, dated 24 December 2016</td>
</tr>
<tr>
<td>swh:1:snp:c7c108084bc0bf3d81436bf980b46e98bd338453</td>
<td>a snapshot of the entire Darktable repository (4 May 2017, GitHub)</td>
</tr>
</tbody>
</table>

Current resolvers: [archive.softwareheritage.org](http://archive.softwareheritage.org) and [n2t.org](http://n2t.org)
Outline

1. The Software Heritage initiative
2. Looking for the right PIDs
3. Demo time
4. Conclusion
Identifiers in action

http://archive.softwareheritage.org/browse
Outline

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
IDOAs 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?