Software Citation WG
Context and first outcomes

P Alliez, R Di Cosmo, B Guedj, A Girault, M-S Hacid, A Legrand, N Rougier

roberto@dicosmo.org

May 28th, 2019
1. Software Source Code: a (forgotten) pillar of Science
2. Software Citation Working Group
3. Proposals
4. The road ahead
Software is everywhere in modern research

[...] software [...] essential in their fields.

Top 100 papers (Nature, 2014)

Sometimes, if you don’t have the software, you don’t have the data

Christine Borgman, Paris, 2018

Nota bene

The links in the picture are essential
Where we stand

Lack of recognition
not (yet) a first class citizen
- in the EOSC plan
- in the scholarly works

Lack of proper guidance on how to
- *archive* software
- choose a license
- *cite* a software project

... but a wealth of initiatives!
- Policies: ACM Artifact Review and Badging, …
- Working groups: FORCE11, RDA, SPSO, …
- Metrics: Open Science Monitor (Elsevier!), …
- Repositories: FigShare, Zenodo, …
Pressure to make the source code available is raising

Why

Necessary to

- *reproduce* and verify,
- *modify* and *evolve*, building new experiments from old ones

When and where

- debate started end of first 2000 decade (biology, statistics, medicine, etc.)
- growing in Computer Science since the ESEC/FSE 2011 Artifact Evaluation context (winner: Vouillon and Di Cosmo)
What is at stake

**Metadata**
Research software artifacts must be properly **described**
make it easy to **discover** them (**visibility**)

**Archival**
Research software artifacts must be properly **archived**
make it sure we can **retrieve** them (**reproducibility**)

**Identification**
Research software artifacts must be properly **referenced**
make it sure we can **identify** them (**reproducibility**)

**Citation**
Research software artifacts must be properly **cited** *(not the same as referenced)*
to give **credit** to authors (**evaluation!**)
1. Software Source Code: a (forgotten) pillar of Science

2. Software Citation Working Group

3. Proposals

4. The road ahead
The Software Citation WG at Inria

Mission
- understand the landscape
- collect best practices
- identify potential Inria contributions
- make recommendations

First outcome
A position paper (submitted to CiSE), available at https://hal.archives-ouvertes.fr/hal-02135891

Here are the key highlights
## Literature review on research software

**Large number of articles**
- reproducibility
- preservation
- crediting developers
- principles/guidelines

**Most common limitations**
- software as ‘just data’
- reduce credit to DOIs
- conflate citation and reference
- citation produced by automated tools

**A few remarkable exceptions**
- **ASCL** (since 1999): metadata only, carefully curated
- **geodynamics.org**: source, documentation, metadata
- **swmath.org**: software catalog via articles
Why it is not simple

Software is complex

- **Structure**: monolithic/composite; self-contained/external dependencies
- **Lifetime**: one-shot/long term
- **Community**: one man/one team/distributed community
- **Authorship**: complex set of roles *(more later)*
- **Authority**: institutions/organizations/communities/single person

Various granularities

- **Exact status of the source code** for reproducibility, e.g.
  
  "you can find at swh:1:cnt:cdf19c4487c43c76f3612557d4dc61f9131790a4;lines=146-187 the core algorithm used in this article"

- **(Major) release**: "This functionality is available in OCaml version 4"

- **Project**: "Inria has created OCaml and Scikit-Learn".
Introspection of Inria best practices

Three places for software

- **CE** evaluation of researchers/teams
- **APP** tech transfert
- **BIL** internal catalog

What we found

- ontologies
- processes
- tools
Outline

1. Software Source Code: a (forgotten) pillar of Science
2. Software Citation Working Group
3. Proposals
4. The road ahead
Three proposals for the scholarly world

A refined ontology for contributors

- Design, Architecture,
- Coding, Testing, Debugging,
- Documentation, Maintenance, Support,
- Management

Similar ideas in CRediT, Geodynamics

We need the human in the loop

When *credit* is at stake, automation/crowdsourcing is not enough; humans intervention is needed to get *quality information*

Reference is distinct from citation

*Reference* is for *reproducibility*, *Citation* is for *credit*: they must not be conflated.
Inria’s contribution to the future of Science

Software Heritage
universal archive (research) software source code archived and referenced

Reproducibility
- tools Guix (now with Software Heritage)
- training/research RR workshops, MOOC

Research software curation
HAL - SWH bridge curation of metadata, and deposit in Software Heritage
Outline

1. Software Source Code: a (forgotten) pillar of Science
2. Software Citation Working Group
3. Proposals
4. The road ahead
Citation for software

- complete ontology
- formal process
- guidelines

Reproducibility

- complete toolchain
- experimentation
- guidelines

Adoption

- conversation started in various communities; we are not really part of it
- simplistic recommendations and tools are being rolled out
- directives are coming soon
- we must take part in the conversation, and disseminate our views

We need open tools supporting the process

Look at Software Heritage, Guix, HAL, BIL, etc. as starting blocks (to align)
**Conclusion**

Research software
- not just data
- finally in the limelight
- soon part of *research evaluation*

Doing it right is not easy
- join the conversation
- *contribute our experience*
- make it into *best practices*

Thank you!
Appendix
5. More on source code for research artefacts

6. Overview of Software Heritage

7. Under the hood

8. Identifiers are not easy

9. Looking for the right identifiers
Source code is special

**Executable and human readable knowledge**

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

Harold Abelson

**Software evolves over time**

- projects may last decades
- the development history is key to its understanding

**Complexity**

- *millions* of lines of code
- large web of dependencies
  - easy to break, difficult to maintain
- sophisticated *developer communities*
ACM take on Reproducibility, Replicability and Source code

ACM policies: Artifact Review and Badging

Terminology (not consensual yet!)

- **Repeatability** \( \text{same team, same experimental setup} \)
- **Replicability** \( \text{different team, same experimental setup} \)
- **Reproducibility** \( \text{different team, different experimental setup} \)

Badging software artefacts

- Functional
- Reusable
- Available
- Results replicated
- Results reproduced

(CC-BY 4.0) Context and first outcomes May 28th, 2019
The state of the art is not ideal

Analysis of 613 papers

- 8 ACM conferences: ASPLOS’12, CCS’12, OOPSLA’12, OSDI’12, PLDI’12, SIGMOD’12, SOSP’11, VLDB’12
- 5 journals: TACO’9, TISSEC’15, TOCS’30, TODS’37, TOPLAS’34

all very practical oriented

The basic question

can we get the code to build and run?
That’s a whopping 40% of non reproducible works!

The main reasons

source code (or the right version of it) cannot be found
Outline

5. More on source code for research artefacts
6. Overview of Software Heritage
7. Under the hood
8. Identifiers are not easy
9. Looking for the right identifiers
# Software Heritage

**Mission**

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

<table>
<thead>
<tr>
<th>Partners</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initiator</strong></td>
</tr>
<tr>
<td><strong>Industry philanthropic sponsors</strong></td>
</tr>
<tr>
<td><strong>Public sector</strong></td>
</tr>
</tbody>
</table>
A revolutionary infrastructure for research and innovation

Reference archive for research software

- curated deposit of research software
- prototype with HAL, CCSD and Inria IES
- 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...
The largest software source code archive ever

10 billions intrinsic identifiers for reproducibility
See DIO vs IDO in bit.ly/swhpipdpaper

Reference archive
See the work done at swmath.org

SWH IDs now a standard for Wikidata
See https://www.wikidata.org/wiki/Property:P6138

Policy
Now part of the French National Plan for Open Science
5. More on source code for research artefacts
6. Overview of Software Heritage
7. Under the hood
8. Identifiers are not easy
9. Looking for the right identifiers
Coverage

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!
Outline

5. More on source code for research artefacts
6. Overview of Software Heritage
7. Under the hood
8. Identifiers are not easy
9. Looking for the right identifiers
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


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.


Scholar roster of broken links

An example from Astronomy

<table>
<thead>
<tr>
<th>Domain</th>
<th>links (broken)</th>
<th>.html</th>
<th>.txt</th>
<th>.dat</th>
<th>.gz</th>
<th>.tar</th>
<th>.fts</th>
<th>tilde</th>
</tr>
</thead>
<tbody>
<tr>
<td>cfa.harvard.edu</td>
<td>802 (110)</td>
<td>336 (70)</td>
<td>0</td>
<td>4 (2)</td>
<td>5 (4)</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>hessarc.gsfc.nasa.gov</td>
<td>640 (33)</td>
<td>423 (27)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.stsci.edu">www.stsci.edu</a></td>
<td>498 (61)</td>
<td>205 (26)</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15 (10)</td>
<td></td>
</tr>
<tr>
<td>asc.harvard.edu</td>
<td>471 (152)</td>
<td>212 (99)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (1)</td>
<td></td>
</tr>
<tr>
<td>cxc.sps.caltech.edu</td>
<td>427 (194)</td>
<td>135 (70)</td>
<td>3 (3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>cfa-www.harvard.edu</td>
<td>332 (68)</td>
<td>277 (52)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>54 (17)</td>
<td></td>
</tr>
<tr>
<td>archive.stsci.edu</td>
<td>308 (58)</td>
<td>57 (9)</td>
<td>2</td>
<td>1 (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.ipac.caltech.edu">www.ipac.caltech.edu</a></td>
<td>285 (14)</td>
<td>209 (12)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.stsci.edu">www.stsci.edu</a></td>
<td>211 (21)</td>
<td>12 (6)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7 (5)</td>
<td></td>
</tr>
<tr>
<td>space.mit.edu</td>
<td>193 (10)</td>
<td>58 (5)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2 (1)</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.astropixi.edu">www.astropixi.edu</a></td>
<td>186 (6)</td>
<td>103 (1)</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.eso.org">www.eso.org</a></td>
<td>186 (58)</td>
<td>54 (22)</td>
<td>1 (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4 (1)</td>
<td></td>
</tr>
<tr>
<td>lisa.ipac.caltech.edu</td>
<td>163 (5)</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.sdss.org">www.sdss.org</a></td>
<td>156 (2)</td>
<td>106 (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>hea-www.harvard.edu</td>
<td>125 (37)</td>
<td>42 (17)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>26 (16)</td>
</tr>
<tr>
<td>physics.nist.gov</td>
<td>125 (3)</td>
<td>63 (2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.nasa.gov">www.nasa.gov</a></td>
<td>120 (3)</td>
<td>50 (2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.nasa.gov">www.nasa.gov</a></td>
<td>120 (3)</td>
<td>50 (2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.hubble.esa.es">www.hubble.esa.es</a></td>
<td>116 (35)</td>
<td>23 (19)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (1)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.astroph.princeton.edu">www.astroph.princeton.edu</a></td>
<td>115 (31)</td>
<td>43 (14)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>53 (12)</td>
<td></td>
</tr>
<tr>
<td>ad.usno.navy.mil</td>
<td>110 (27)</td>
<td>98 (22)</td>
<td>3 (3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (1)</td>
<td></td>
</tr>
</tbody>
</table>

This table lists total number of links and broken links (HTTP status codes 300, 400, and 500) 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.0104798.000

How Do Astronomers Share Data?
Pepe, Goodman, Muench, Crosas, Erdmann

PLOS August 28, 2014
dx.doi.org/10.1371/journal.pone.0104798
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*
Outline

5. More on source code for research artefacts
6. Overview of Software Heritage
7. Under the hood
8. Identifiers are not easy
9. Looking for the right identifiers
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 (minting)</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
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 make restrictions that prevent others from denying you these rights or asking you to surrender the rights. These restrictions take the form of an easement from all property on that copyright. These restrictions make sure that the copyright holder cannot阻止 you from working with the program as a whole. For details, see Sections 1 and 2.

However, there is nothing stopping you from making a free program into a non-free program, as long as you change it and thus choose not to use it in the combined work as a whole, i.e. to make an essential part of the removed from it.

The precise terms and conditions for copying, distribution and modification follow.

0. Definitions

"License" refers to this license document.
"Legal Requirements" refers to society's norms about copyright and limitations on copying or sharing.
"Program" refers to a basic unit of source code by which the License applies, and all the source and object files to which it is applicable.
"Copy" means复制.
"Convey" means either distribution or public performance.r.
"Licensees" means anyone who is granted a license by this License, either originally or through a conveyance of a copy from a licensor. "Recipients" means Licensees and all their downstream recipients.

1. Source Code

The "Source Code" for a Program is the preferred form in which to make available or receive copies of the Program, including new versions, modified versions, and additions of existing versions. Source code means the source code form of a Program not modified or changed except by automatic means defined above. Any operation, other than an operation covered by Sections 3 through 7, is allowed for non-source code.

2. Grant of Copyright License

This License grants a copyright license for the<br>Program. Under this License, no one reserves or retains any ownership right in the Program. Instead, under this License, the copyright holder grants you permission to reproduce, distribute, prepare derivative works, and perform the Program as specified below.

3. Redistribution

You may convey copies of the Program or modifications thereof to any person to whom you provide any of the following documents for the Program:
- The complete corresponding machine-readable source code, which must be freely distributed under this License (including with any changes you made in other parts of the source code that you own or control), and<br>- Any offer to distribute the source code for the Program under the terms of this License (including any changes you made in other parts of the source code that you own or control).
- The complete machine-readable object code of a work derived from the Program with the Franctions described in Section 4, if the work has the kind of code described in Section 4.
- A written offer, timed to expire in (say) one year after the date of distribution, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the source code of the<br>whole work, in the same format as you received it, so they can make copies of the modified version as specified in Section 3. If you ask to keep the name or any portion of the name of the original author, you must do so, in the case of source code, by putting the name clearly in the machine-readable text shows that you changed it, and giving the name to the recipients in the files when they are released under these conditions.

You may convey a work based on the Program under the terms of Sections 3 through 7, provided that you accompany it with the machine-readable object code of the work under the terms of Sections 3 through 7. You may convey one or more works based on the Program (or a modified version of it) under the terms of Sections 3 through 7, provided that you offer equivalent交错的 waters, and that the equivalent交错的 waters are available under the terms of this License. In addition, you may copy the work or works in unmodified object code form and convey it without copying the source code, under the terms of this License, provided that you agree to embed a notice which states that you did not modify the program and that the Program is covered by this License.

You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any other attempt to copy, modify, sublicense, or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in this license agreement. This License does not deny the rights of any other parties with the rights expressed in this License.

The Free Software Foundation may publish new versions of this License from time to time. You may use any version of this License, created or modified since the announcement of the Free Software Foundation, but only if you explicitly state on the inclusion of the notice that you do so.

4. Special Exceptions

The Free Software Foundation may also publish additional documents that are consistent with this License. The Free Software Foundation may also apply a different license to a specific document. If this document, or any part of it, is so dated, any license expressed in this document (including Sections 3, 4, and 7) will be considered to be in the process of transition to the new license. References to "current" or "existing" versions of this License will be eligible to use the new letter of the new license. This License will apply to any other versions of this License not yet published.

5. Trademarks

This License does not grant permission to use "Linux" in your program's name, nor does it affect the trademark status of words related to "Linux". Any use of "Linux" in your program's name, or in your program itself, is not covered by this License. Such use may violate the laws of certain countries, and does not confer any rights under this License.

The free software movement recognizes the importance of using a standard, recognized name in the free software community. We therefore request that you do not use any name other than the name "Linux" in your program's name, and that you avoid any name that closely resembles the name "Linux" without prior permission from the Free Software Foundation.

6. No凰凰

Nothing in this License shall be construed as granting a license for any patent claim by the Free Software Foundation, either now or in the future. Nothing in this License grants a license for any trademark of the Free Software Foundation, either now or in the future.

7. Disclaimer of Warranty

Unless required by applicable law or agreed to in writing, software distributed under this License is provided "as is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose, and non-infringement. Any later version of this license is included in any later version of the software, except to the extent that subsequent versions of the software provide warranties or conditions that disagree with those provided by this license.

8. Termination

This License will end if you fail to comply with its terms. This license agreement is terminated immediately if you fail to comply with any of the terms of this license agreement.

9. Acceptance

Any use of this license agreement indicates your acceptance of the terms of this license agreement. Any use of this license agreement indicates acceptance of the terms of this license agreement.

10. Adaptation

You may create and convey adaptations of the Program under this license agreement, provided that you name the author of the original work clearly and distinctly.

11. End

This license agreement is terminated immediately if you fail to comply with any of the terms of this license agreement.

End of GNU General Public License
IDOs in Software Heritage: a worked example
Directories

100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d6a34af6f52cf3cf66b0c2f7bd0648fbdb255e7f AUTHORS
100644 blob 94a9ed02d3859793618152aa559a168bcb85e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e8867751dfb95c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5eaf2eb4 Makefile
100644 blob 5c7e3a5b0bddd0b38682ba7793f440492ed9678bb3 Makefile.local
100644 blob 8617980629cd24e6080404f09aa749b6853e677 README.db_testing
100644 blob 76b29f94cf815e0869c414d38d7b7ce08ec514e README.dev
040000 tree ee1e0ceef948af0b93ad8372af89f12e92618a bin
040000 tree 83e56d8beaf7793c77a45a435c80fcb8af503013 debian
040000 tree a34c9c4ba213f9cedc67f9816348d2795557af5 docs
100644 blob f2a6d32c6135aa7287bbd76167b01df2ae4f1539 requirements.txt
100755 blob eee147c36caflbbcc2d820da8dc026c5b68180c setup.py
040000 tree 224bb4c1f4c67fca1d160bff9d2d06094e7e1abf3 sql
040000 tree 8631c9cd77bb993168107ab5ba51f1f40c6300be swh
040000 tree 8fb905b56ba8ed692f1209b2773b474c61d66c1 utils

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
IDOs in Software Heritage: a worked example
Revisions

<table>
<thead>
<tr>
<th>Details</th>
<th>Changes</th>
<th>Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author: Nicolas Dandrimont <a href="mailto:nico@dandrimont.eu">nico@dandrimont.eu</a> (Thu Sep 14 26:13 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Committer: Nicolas Dandrimont <a href="mailto:nico@dandrimont.eu">nico@dandrimont.eu</a> (Thu Sep 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: fc3a8b59ca1df424d860f2c29ab07f8ee4dc35d10</td>
<td>test_storage: properly pipeline origin and cont...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>provenance.tasks: add the revision -&gt; origin cache task</td>
<td></td>
</tr>
</tbody>
</table>

```
  swh/storage/provenance/tasks.py  77
```

tree 515f00d44e92c65322aaa9bf3fa097c00dddb9c7d
parent fc3a8b59ca1df424d860f2c29ab07f8ee4dc35d10
author Nicolas Dandrimont <nico@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nico@dandrimont.eu> 1472732773 +0200

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

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6
IDOs in Software Heritage: a worked example
Releases

object c0c9f16b1e134f593e7567570a1761b156e6eb1d

type commit

tag v0.0.51
tagger Nicolas Dandrimont <nicolas@dandrimont.eu>

Date: Wed Aug 24 14:36:03 2026 +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-----

IQIQIBAAOAAdBiQxZTNfxxuaWVsaGFQGHRnbmRyaW1ibmQuZXMASGkQ7AYLMo2+n neqqiw/aqD50sb5jL0xO+KWN13XqV5+iK1x1EvGlzr/fWaxbEK7x2KXzEIuD7Tuf ahpZ6z3q9Kn6aC1+Yx8fpk3hJL2Ytq3ZeWXXWqrknWMMAnEoTfDBqasphwhhBAD5t2 ICBI2uJXuFLDFq93eKFPwzzZXYg+hBq5MyW3y5Drw6757ZK4JzJPGgyyHP75yo IGEndWno7y/1Y1Vmb12n35b175sMRqAa+becqCQsujT2Z2xj+yjUqUCbyqkN31mr/fL qs72z1z0y8H8WcJZ1qjv+sIlS8m8nPho3F8Ho0jZjyPl9fK/KQCkao k/j6IkAyWU0Meb+nK/V9j/eR3+yYWFbJ30q5a12K/On7h61dALChMFeAk6kO5h4 Mq/3tJ1Zg0E0fmsW67G6sDwK6PK4hflhVLOf0V39G9qvQ7ru1qLmzo606h9fAAbwZ CqGkY1PdhT4v59N1id4wYFP2ywe0U2XKDF3dVWvF2Z1Wm+52tMzckDr8y3U0O0m RpFTTUSxXUeHXGOpkgkX5YTrw1pQ7C6USRTkOD6G48Azm1k0gGmGyXCVfpyPb
mh1bSBHBNMoqF6y75oqub1K73p1FRUKGWdWEKgk5sxWKUZGK5y67jQ1329

gu9qZm9fQ25t806oKAL2+Hv9PaBlucKmPu6g6zi3+EHyvUK= = KOxP

----END PGP SIGNATURE-----
IDOs in Software Heritage: a worked example
Snapshots

commit 00ffeb2577010852eb3c21691466c53a1d9158 refs/heads/atime
commit ba5443a24ae9f4e32ba46c292ce4f4c2e61c207eb refs/heads/directory-listing-arrays
commit df896da892833f5689b9cb13d562723869c5 refs/heads/proprietary
commit c7ff9e6e9e62b27806a3f075a8610f67de60ee888 refs/heads/master
commit 7ecac197fc6d5ed2b2a047e54b1ed9e8b443616a8fe2 refs/heads/tmp-dir-release
commit e442d285f7105895a565027b53ee4f02522e62e refs/heads/tmp-release
commit b910f43b137b678096659779f46907c757f755 refs/tags/v0.3

tag 72a2191a338a5e39996d08b7080b07ee2acfed refs/tags/v0.4.0

tag 35960e8c48eb785d37e6bfa23b8bbf4f4f50c refs/tags/v0.5.1

tag 3337842e7a33baa5e97777b8e0b66740c535a refs/tags/v0.6.1

tag 86f7465275232cf5603b11c2bfa36c34b3ab5d refs/tags/v0.6.2

tag 5a6352f5e860d4e4b566e7d92a1e3273b7b refs/tags/v0.6.3

tag 58e6fa46a58b4f55c8b57f5907643cb03a9cf77 refs/tags/v0.6.4

tag 8c0d8885f4f98b363177742b289f660e50b51c refs/tags/v0.6.5

tag a54244ae3f98be5eb4262ee83200936763d8 refs/tags/v0.6.6

tag 22baf23559d1222265555462e1e88054993d9 refs/tags/v0.6.7

tag 669979a4c054d977fc0d24aad0d0c82536e4f7c refs/tags/v0.6.8

tag 33b2f1359f1d222265555462e1754993d9 refs/tags/v0.6.9

tag 3148c3d3eac4c6c76d279801e94e0b1237edf72c7 refs/tags/v0.6.10

tag 215ea5080a11108e20d072766e0b70783a78900 refs/tags/v0.6.11

tag 3fb6d8c275a5d5d252124257a5d5dfc855fa1df refs/tags/v0.6.12

tag 8c0bee68da4d73f5e4267898e4b16ac3c7e2ab4a refs/tags/v0.6.13...
The Software Heritage IDO schema (see http://bit.ly/swphpids)

<table>
<thead>
<tr>
<th>ID</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>swh:1:cnt</td>
<td>94a9ed024d3859793618152ea559a168bbcbbb5e2</td>
<td>full text of the GPL3 license</td>
</tr>
<tr>
<td>swh:1:dir</td>
<td>d198bc9d7a6bcf6db04f476d29314f157507d505</td>
<td>Darktable source code</td>
</tr>
<tr>
<td>swh:1:rev</td>
<td>309cf2674ee7a0749978cf8265ab91a60aea0f7d</td>
<td>a revision in the development history of Darktable</td>
</tr>
<tr>
<td>swh:1:rel</td>
<td>22ece559cc7cc2364edc5e5593d63ae8bd229f9f</td>
<td>release 2.3.0 of Darktable, dated 24 December 2016</td>
</tr>
<tr>
<td>swh:1:snp</td>
<td>c7c108084bc0bf3d81436bf980b46e98bd338453</td>
<td>a snapshot of the entire Darktable repository (4 May 2017, GitHub)</td>
</tr>
</tbody>
</table>

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