Citing 65 Million Software Projects
Leveraging the Software Heritage archive for software citation

Stefano Zacchirolı
University Paris Diderot & Inria
zack@upsilon.cc

October 26, 2017
The Software Heritage Project

Our mission

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

Past, present and future

Preserving the past, enhancing the present, preparing the future.
Our principles

- Open approach
  - open source
  - transparency

- In for the long haul
  - non profit
  - replication

Software Heritage
Data flow

Software Heritage Archive

Forges

GitHub lister

GitLab lister

Debian lister

PyPi lister

Distros

Package repos

Listing (full/incremental)

Loading & deduplication

Scheduling

Distros

...
Our current sources

- GitHub
- Debian, GNU
- WIP: Gitorious, Google Code, Bitbucket
Archive coverage

Our current sources

- GitHub
- Debian, GNU
- WIP: Gitorious, Google Code, Bitbucket

150 TB blobs, 5 TB database (as a graph: 7 B nodes + 60 B edges)
Archive coverage

Our current sources

- GitHub
- Debian, GNU
- WIP: Gitorious, Google Code, Bitbucket

150 TB blobs, 5 TB database (as a graph: 7 B nodes + 60 B edges)

The richest source code archive already, … and growing daily!
## Our requirements in the PID arena

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long term</td>
<td>Identifiers must be there for the long term</td>
</tr>
<tr>
<td>No middle man</td>
<td>Identifiers must be meaningful even if resolvers go away</td>
</tr>
<tr>
<td>Integrity, not just naming</td>
<td>Identifier must ensure that the retrieved object is the intended one</td>
</tr>
<tr>
<td>Uniqueness by design</td>
<td>only one name for each object, each object has only one name</td>
</tr>
</tbody>
</table>
Intrinsic identifiers in Software Heritage

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

Combination of
- tree
- hash function

Stefano Zacchioli
Intrinsic identifiers in Software Heritage

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
Back to basics: DIOs vs. IDOs

DIO (digital identifier of an object)

- digital identifiers for traditional (non digital) objects
  - epistemic complications (manifestations, versions, locations, etc.)
  - significant governance issues, …
Back to basics: DIOs vs. IDOs

DIO (digital identifier of an object)
- digital identifiers for traditional (non digital) objects
  - epistemic complications (manifestations, versions, locations, etc.)
  - significant governance issues, …

IDO (identifier of a digital object)
- (digital) identifier for digital objects
  - much simpler to build/handle
  - can (and must) be intrinsic
Back to basics: DIOs vs. IDOs

**DIO (digital identifier of an object)**
- digital identifiers for traditional (non digital) objects
  - epistemic complications (manifestations, versions, locations, etc.)
  - significant governance issues, ...

**IDO (identifier of a digital object)**
- (digital) identifier for digital objects
  - much simpler to build/handle
  - can (and must) be intrinsic

**Separation of concerns**
- we need *both* DIOs and IDOs
- we should not mistake DIOs for IDOs (and viceversa)
Referencing archived software

Use case #1 — reference archived source code artifacts
Use case #1 — reference archived source code artifacts

directory \texttt{swh:1:dir:06741c8c37c5a384083082b99f4c5ad94cd0cd1f}
id of tree object listing all the files in a project (at a given time)
Referencing archived software

Use case #1 — reference archived source code artifacts

- **directory** `swh:1:dir:06741c8c37c5a384083082b99f4c5ad94cd0cd1f`
  id of tree object listing all the files in a project (at a given time)

- **revision** `swh:1:rev:7598fb94d59178d65bd8d2892c19356290f5d4e3`
  id of commit object which a tree and (a pointer to) the history
Referencing archived software

Use case #1 — reference archived source code *artifacts*

**directory**  
`swh:1:dir:06741c8c37c5a384083082b99f4c5ad94cd0cd1f`  
id of tree object listing all the files in a project (at a given time)

**revision**  
`swh:1:rev:7598fb94d59178d65bd8d2892c19356290f5d4e3`  
id of commit object which a tree and (a pointer to) the history

Use case #2 — reference an abstract software *product / "project"*

**metadata**  
this *will* involve some form of DIO (and we get all the complications back)
A first prototype?
Daniel S. Katz, *Software Heritage and repository metadata: a software citation solution*,

Come in, we’re open
www.softwareheritage.org

Roberto Di Cosmo, Stefano Zacchirolı
*Software Heritage: Why and How to Preserve Software Source Code*
The archive in a few pictures

A giant (extended) Merkle DAG
The archive in a few pictures

A giant (extended) Merkle DAG
The archive in a few pictures

A giant (extended) Merkle DAG

Directories

- .gitignore
- AUTHORS
- LICENSE
- MANIFEST.in
- Makefile
- Makefile.local
- README.db_testing
- README.dev
  - bin
  - debian
  - docs
  - requirements.txt
  - setup.py
  - sql
  - swh
  - utils

```
100644 blob c5baade4c44766042186ef858c8f63d587efb09 .gitignore
100644 blob 2d0a34af6f52cf3c68c2f7bd0648fbdb255e77f AUTHORS
100644 blob 949e6d824d3859793618152ea559a168bbcbb25e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e867751dfb095c7bb MANIFEST.in
100644 blob 524175c2badb35b975f79284c2f5a6d5eaf2eb4 Makefile
100644 blob 5c7e3a5bbdb083682ba7793f440492ed9678bb3 Makefile.local
100644 blob 8617986629c2d4e688040f19aa749b885b3e87b README.db_testing
100644 blob 76b29f94cf815e0869c414d38d78d139e98ec514e README.dev
040000 tree ele10cecf948af08093ad0372af689f12e92618a bin
040000 tree 83e56d0bea7793c77a45a3a45c80f1cb8af583013 debian
040000 tree a34c9c4ba213f9c6cd867f9816348d27955779af5 docs
100644 blob f2a6d32c6135aa728bbd7617b091df2af4f1539 requirements.txt
100755 blob eee1c473c3afa1bbcc2d820da8dc026cb5b68180bc setup.py
040000 tree 224bb41cf4c67fca1d160b0f8d2d6e091ae7e1abf3 sql
040000 tree 8631c9c777be9993168107ab5baf51f40c63000e swh
040000 tree 8f905b56ba8ed692f1209b2773b474c61d66c1 utils
```

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d

Stefano Zacchirol

Citing 65 Million Software Projects October 26, 2017
The archive in a few pictures

A giant (extended) Merkle DAG

Snapshots → Releases → Revisions → Directories → Contents
The archive in a few pictures

A giant (extended) Merkle DAG

Revisions

- **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

```
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 a few pictures

A giant (extended) Merkle DAG
The archive in a few pictures

A giant (extended) Merkle DAG

Releases

object c0c9f16b1e134f593ed756757b1761b156e6eb1d

commit c0c9f16b1e134f593ed756757b1761b156e6eb1d

A giant (extended) Merkle DAG

Releases

object c0c9f16b1e134f593ed756757b1761b156e6eb1d

type commit

tag v0.5.1

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

- Add new metadata column to origin_visit
- Update swh-add-directory script for updated API

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

iQz8AAABCAAdEIQXz7FPluxWVygG2QG4phbmnRyAVWVzbnQjZXXUACgkJ7A7WLo2a
nephraJqQgS0bU1p2z7lg4wiOUsiKpGyS=1KjVv3hJ1wKXawxE4q7aX2KxL2HT3uf
ahpZ6qS3qR8os6AC1+YXK1fC3L2YysZ6XWQp4XwVMA5e8vDBbqanwLH4D5s2
IC83v2gXxU/0X9ixKWWzwZ2Sg=+lOnMV35YR6Lw7Z3YK4jWcLp3g3yH75Y5Sv
igXEOw0mY7H1Lvm61MoLU27XnRkap+beq3GhJaJZgty=+iYj7Lj3Cv2qLmlh
q42m98yylxhV1G1vJXfY7Sd0ST+7I4bOYjFPOXhjKpFS79QvQH2ZfPKzCaO
k+Jlw4KqL8QMo6-+xXVjxL9x3+-YXHJ3>GpsaJlX6b07?n6iJlAdUcEoXvXKxHtM;
d/0gMxR4kU1+kpy WebbW75G5sbW4wKkPHygVQ3vJ3Qv3GqQtu7Rq80O6Hw6wzC
GgKJ1478JH59H56Hw4FZyzy+Q2VXOLs6WvLV94q2hK42W5J8=+O2mbGRePE0jSUOMn
Rq8Fb8xUKXVdG020qJ5h7vno1gRGC1JU5O7k5aGHB4Mz1UKmGyN0wKCVqPoYo
rhbB9WqBNH7aoyF6TsQunbYK7QyYRUGKWRdRKe0xSxxHKJ5ZGTCZy6Y9jl029
qz/0y7Gq5WQCO8oonzA12+HvFw+sYcH7eUjUng6z2spF=EnvUCm=
s=EOqP

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

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb
The archive in a few pictures

A giant (extended) Merkle DAG