Software Heritage
Une Archive Mondiale du Logiciel Libre, Inspirée de Git

Nicolas Dandrimont, Stefano Zacchirolí

September 7th, 2016
Meetup Git
Paris, France
1. The need for software preservation
   - Software all around us
   - Software is fragile

2. The Software Heritage project
   - Mission
   - Status and roadmap

3. Software Heritage internals
   - Archive
   - Git loader

4. Conclusion
Software is Pervasive

At the heart of our society

- communication, entertainment
- administration, finance
- health, energy, transportation
- education, research, politics
- ...

N. Dandrimont, S. Zacchioli
Software is Pervasive

At the heart of our society

- communication, entertainment
- administration, finance
- health, energy, transportation
- education, research, politics
- ...

At the heart of technology

- house appliances ≈ 10M SLOC
- phones ≈ 20M SLOC, cars ≈ 100M SLOC
- Internet of things, …
Key mediator for accessing all information (c) Banksy

Information is a main pillar of our modern societies.

Absent an ability to correctly interpret digital information, we are left with [...] "rotting bits" [...] of no value.

Vinton G. Cerf IEEE 2011
Information is a main pillar of our modern societies. Absent an ability to correctly interpret digital information, we are left with […] "rotting bits" […] of no value.

Vinton G. Cerf IEEE 2011

[…] the vast majority describe experimental methods or software that have become essential in their fields.

Top 100 papers (Nature, October 2014)
Software is Knowledge

Key mediator for accessing all information (c) Banksy

Information is a main pillar of our modern societies.

Absent an ability to correctly interpret digital information, we are left with […] "rotting bits" […] of no value.

Vinton G. Cerf IEEE 2011

Software is an essential component of modern scientific research

[…] the vast majority describe experimental methods or software that have become essential in their fields.

Top 100 papers (Nature, October 2014)

Bottomline: Software embodies our Knowledge and Cultural Heritage

It must be collected, preserved, referenced and made accessible!
Software is Fragile

Bits rot, hosters shut down
- Have you tested your backups recently? How about git fsck?
- Gitorious
- Google Code

Software is scattered all around
GitHub, GitLab, BitBucket, SourceForge, alioth, …
… your personal home page, …

No uniformity or stability whatsoever
Software migrates from hosters to hosters, URIs aren’t perennial
1. The need for software preservation
   - Software all around us
   - Software is fragile

2. The Software Heritage project
   - Mission
   - Status and roadmap

3. Software Heritage internals
   - Archive
   - Git loader

4. Conclusion
Our mission

Software Heritage

Collect, organise, preserve and share all the software source code that lies at the heart of our culture and our society.

https://www.softwareheritage.org/
Software Source Code is different

“Programs must be written for people to read, and only incidentally for machines to execute.” Harold Abelson, Structure and Interpretation of Computer Programs

Distinguishing features

- executable and human readable knowledge (an all time new)
  - even hardware is… software! (VHDL, FPGA, …)
  - text files are forever
- naturally evolves over time
  - the development history is key to its understanding
- complex: large web of dependencies, millions of SLOCs

In a word

- software is not just another sequence of bits
- a software archive is not just another digital archive
Current status

Ingest all the software

- all (non-fork) GitHub repositories
- all Debian package uploads from snapshot.debian.org
- the GNU project FTP archive

Preserve all the software

- Google Code
- Gitorious
Features

- *lookup* by hashes for contents (done)
- *browse*: wayback machine for software source code
- *download*: git clone from Software Heritage
- *provenance* information for all the content
- *full text search*: dive into the Software Heritage archive
## Roadmap

### Features

- *lookup* by hashes for contents (done)
- *browse*: wayback machine for software source code
- *download*: git clone from Software Heritage
- *provenance* information for all the content
- *full text search*: dive into the Software Heritage archive

### ... and much more than one could imagine

- license information (for code reuse)
- source code autocomplete in your IDE
- FOSS history/archeology
- the world’s software devel history in a single Merkle DAG!
Outline

1. The need for software preservation
   - Software all around us
   - Software is fragile

2. The Software Heritage project
   - Mission
   - Status and roadmap

3. Software Heritage internals
   - Archive
   - Git loader

4. Conclusion
The structure of the archive

On-disk storage
- flat file storage for contents
- postgres database for the metadata

Data model: one big Merkle DAG, inspired by the git model
- Origins (= repositories)
- Snapshots (= lists of branches and tags)
- Releases (= tags)
- Revisions (= commits)
- Directories (= trees)
- Contents (= blobs)
The archive in pictures

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 or asking you to surrender the rights. Therefore, you receive
a license to do what the license authorizes. There can be no legal
extent to that authorization other than what the license in a particular
file grants. If you distribute copies, youPixels
The archive in pictures
### Directories

<table>
<thead>
<tr>
<th>ID</th>
<th>Blob Size</th>
<th>Hash</th>
<th>Filename</th>
</tr>
</thead>
<tbody>
<tr>
<td>100644</td>
<td>20380</td>
<td>1c5baade4c44766042186ef858c0fd63d587ebf09</td>
<td>.gitignore</td>
</tr>
<tr>
<td>100644</td>
<td>20380</td>
<td>2d0a34af6f52cf3cf6b0c2f7bd0648fbdb255e77f</td>
<td>AUTHORS</td>
</tr>
<tr>
<td>100644</td>
<td>20380</td>
<td>94a9ed024d3859793618152ea559a168bcb0bb5e2</td>
<td>LICENSE</td>
</tr>
<tr>
<td>100644</td>
<td>20380</td>
<td>d9b2665a435a43f8a79a84e0867751db095c7bb</td>
<td>MANIFEST.in</td>
</tr>
<tr>
<td>100644</td>
<td>20380</td>
<td>524175c2bad0b35b97f579284c2f5a6d5eaf2eb4</td>
<td>Makefile</td>
</tr>
<tr>
<td>100644</td>
<td>20380</td>
<td>5c7e3a5bddd038682ba7793f440492ed9678bb3</td>
<td>Makefile.local</td>
</tr>
<tr>
<td>100644</td>
<td>20380</td>
<td>8617980629cd24e6080404f09a749b085b3e07b</td>
<td>README.db_testing</td>
</tr>
<tr>
<td>100644</td>
<td>20380</td>
<td>76b29f94cf815e0869c414d38d78d7ce08ec514e</td>
<td>README.dev</td>
</tr>
<tr>
<td>040000</td>
<td>11521</td>
<td>e1e10ecef948af0b93adb0372af89f12e92618a</td>
<td>bin</td>
</tr>
<tr>
<td>040000</td>
<td>11521</td>
<td>83e56d0beaf7793c77a45a345c80fcb8af503013</td>
<td>debian</td>
</tr>
<tr>
<td>040000</td>
<td>11521</td>
<td>a34c9c4ba213f0cedc67f9816348d27955577af5</td>
<td>docs</td>
</tr>
<tr>
<td>100644</td>
<td>11521</td>
<td>f2a6d32c6135aa7287bbd76167b01df2ae4f1539</td>
<td>requirements.txt</td>
</tr>
<tr>
<td>100755</td>
<td>11521</td>
<td>eee147c36caflbb2d820da8dc026cb5b68180bc</td>
<td>setup.py</td>
</tr>
<tr>
<td>040000</td>
<td>11521</td>
<td>224bb4c1f4c67fca1d60bff2d06094e7e1abf3</td>
<td>sql</td>
</tr>
<tr>
<td>040000</td>
<td>11521</td>
<td>8631c9c77bbe993168107ab5baf51f40c6300be</td>
<td>swh</td>
</tr>
<tr>
<td>040000</td>
<td>11521</td>
<td>8fb905b56ba8ed692f1209b2773b474c6c1d66c1</td>
<td>utils</td>
</tr>
</tbody>
</table>

id: 515f00d44e92c65322aa9bf3fa097c00ddb9c7d
The archive in pictures

Snapshots → Releases → Revisions → Directories → Contents
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:nicolas@dandrimont.eu">nicolas@dandrimont.eu</a> (Thu Sep 1 14:26:13 2016)</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 2016)</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: fc3a8b59ca1df424d860f2c29ab07fee4dc35d10 : test...storage: properly pipeline origin and cont...provenance.tasks: add the revision -&gt; origin cache task</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

```
515f00d44e92c65322aaa9bf3fa097c00db9c7d
tree
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
The archive in pictures

Releases

object c0c9f16b1e134f593e7567570a1761b156e6eb1d

Release swh.storage v0.51

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

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

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb
The archive in pictures
Snapshots

commit 08ffe2570190525eb3ce21691466c53a1d9158 refs/heads/ptime
commit ba5443a24e3f9fe432a46c292cecc4fcede61c67eb refs/heads/directory-listing-arrays
commit d69eb0dbf892383ff6589b27fbeb1c05d27238d9c5 refs/heads/foo
commit cf7fff9ea0e6b22f8946908f5a8019f67de468e08 refs/heads/master
commit 7ec0a197f3e66d2024047e554b1ed9e8b88d9663610fc2 refs/heads/tmp/directory-add
commit 642a2a05f37de85805a85d427b53ee4fb2252e82e refs/heads/tmp/generic-releases
tag 20f04cb137cf7f69d966597799fd4907c757f755 refs/tags/v0.0.1
tag 72a12991a384e539996d6bb677f0b8bbee72ae2cd refs/tags/v0.0.10
tag 35996ea0eb0b70e5b376705fa23bbfba4fa5cc refs/tags/v0.0.11
tag 3337427a403ba569a67777b8d5f86674fbc6556 refs/tags/v0.0.12
tag 06f74652755b337cf59e331c26a0d60c9f3e735d refs/tags/v0.0.13
tag 5a6325fe86ab854b581d7442667d92a11e32f3bd refs/tags/v0.0.14
tag 586fbae580b4f5faba05f599367643ccbcbaa9c7f refs/tags/v0.0.15
tag 8cd88b8b5f40998b3631777742db289f660e5be51c refs/tags/v0.0.16
tag a542244e3e30fbed35efb2b20fee035809a9bc76 refs/tags/v0.0.17
tag 228a2f1650dd2222e55659462i0e66c4993d9 refs/tags/v0.0.18
tag 6069794ca059d497fc0b242a000dce82636ef4c7c refs/tags/v0.0.19
tag 32b5a59fc2a323baa6d5f15ada0d338b2e75b9a7 refs/tags/v0.0.2
tag 3147c3934c6c6492f881e906b1237ebd9f2c7 refs/tags/v0.0.20
tag 215ea50da911e082eb0b72e6e0b46073a87908 refs/tags/v0.0.21
tag 3fb16a2c0972a5656252124257a02c0e5f02fafa1df refs/tags/v0.0.22
tag 8c5be8da4d73fc5d262789e460a16aac3c72aba4 refs/tags/v0.0.23

...
Software Heritage in numbers

**Volume**
- 120TB used by (gzipped) files on disk
- 3.1TB PostgreSQL database for the metadata

**Counts**
- 2.7 billion files
- 2.2 billion directories
- 600 million revisions
- 12 million people
- 5 million releases

By far, the biggest DVCS tree in existence

N. Dandrimont, S. Zacchirolı

Software Heritage 07/09/2016, Meetup Git Paris
Software Heritage in numbers

Volume
- 120TB used by (gzipped) files on disk
- 3.1TB PostgreSQL database for the metadata

Counts
- 2.7 billion files
- 2.2 billion directories
- 600 million revisions
- 12 million people
- 5 million releases

By far, the biggest DVCS tree in existence
Git loader architecture

- git Server
- swh git importer
- swh database
Git loader architecture

- Git Server
- swh git importer
- swh database

GET info/refs
Git loader architecture
Git loader architecture

GET info/refs

list of remote refs

swh git importer

list of known objects

last snapshot?

swh database
Git loader architecture

```
GET info/refs
list of remote refs
POST git-upload-pack
  (have xxx want yyy)

last snapshot?
list of known objects

git Server

swh git importer

swh database
```
Git loader architecture

Git Server

swh git importer

Inject missing contents

swh database

- blob a5aa0aa
- commit bbbbbb
- tree ccc1cfc
- blob ccccaaa
- tag ddddddd
- blob 123a452
- commit 938
- tree
Git loader architecture

**Diagram:**
- **git Server**
  - blob: a5aa0aa
  - commit: bbbbbbbb
  - tree: ccc1cfc
  - blob: ccccaaa
  - tag: dddddddd
  - blob: 123a452
  - commit: 938
  - tree

- **swh git importer**
  - Inject missing dirs

- **swh database**

**Textual Representation:**
- Git loader architecture
- N. Dandrimont, S. Zacchirol
- So/f_tware Heritage 07/09/2016, Meetup Git Paris 12 / 13
Git loader architecture

- **Git Server**
  - blob: a5aa0aa
  - commit: bbbbbbb
  - tree: ccc1cfc
  - blob: ccccaaa
  - tag: ddddddd
  - blob: 123a452
  - commit: 939
  - tree

- **swh git importer**

- **swh database**

  - Inject missing releases
  - list tags
Git loader architecture

- git Server
- swh git importer
- swh database

Inject new snapshot
1. The need for software preservation
   - Software all around us
   - Software is fragile

2. The Software Heritage project
   - Mission
   - Status and roadmap

3. Software Heritage internals
   - Archive
   - Git loader

4. Conclusion
How to contribute to Software Heritage

Developers

- Info: www.softwareheritage.org/community/developers
- Code: forge.softwareheritage.org

Sponsors

INRIA (founder)

- sponsorship opportunities
  www.softwareheritage.org/support/sponsors

Contact us

- email: {olasd,zack,info}@softwareheritage.org
- twitter: @swheritage
- web: www.softwareheritage.org