Software Heritage: Archive your source code

Benoit Chauvet

Software Heritage
Software Engineering Manager

09 November 2023
Guix HPC Workshop
Montpellier, France

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

Who we are

- Software Heritage
- Started in 2015 (Roberto Di Cosmo & Stefano Zacchiroli)
- Universal archive of source code
- Initiated by Inria, in collaboration with Unesco

Outline

Software is everywhere (and a key mediator) in society

A few basic needs for research software

Archive

Research software artifacts must be properly archived

make sure we can *retrieve* them (*reproducibility*)

A few basic needs for research software

Archive

Research software artifacts must be properly archived

make sure we can *retrieve* them (*reproducibility*)

Reference

Research software artifacts must be properly referenced

make sure we can *identify* them (*reproducibility*)

A few basic needs for research software

Archive

Research software artifacts must be properly archived

make sure we can *retrieve* them (*reproducibility*)

Reference

Research software artifacts must be properly referenced

make sure we can *identify* them (*reproducibility*)

Describe

Research software artifacts must be properly described

make it easy to *discover* and *reuse* them (*visibility*)

A few basic needs for research software

Archive

Research software artifacts must be properly archived

make sure we can *retrieve* them (*reproducibility*)

Reference

Research software artifacts must be properly referenced

make sure we can *identify* them (*reproducibility*)

Describe

Research software artifacts must be properly described

make it easy to *discover* and *reuse* them (*visibility*)

Cite/Credit

Research software artifacts must be properly cited *(not the same as referenced!)*

to give *credit* to authors (*evaluation*!)

Outline

But *where* is this commons?

- many disparate development platforms, with a few dominant players (e.g., GitHub)
- a myriad places where distribution may happen
- most of them operated by for-profit companies

Software source code is fragile

Like all digital information, FOSS is fragile

- link rot: projects are created, moved around, removed
- business-driven code loss (e.g., Gitorious, Google Code, Bitbucket)
- data rot: physical media with legacy software decay

Software source code is fragile

Like all digital information, FOSS is fragile

- link rot: projects are created, moved around, removed
- business-driven code loss (e.g., Gitorious, Google Code, Bitbucket)
- data rot: physical media with legacy software decay

If a website disappears you go to the Internet Archive…

where do you go if (a repository on) GitHub or GitLab goes away?

Outline

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog

find and reference all software source code

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog

find and reference all software source code

Universal archive

preserve and share all software source code

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog

find and reference all software source code

Universal archive

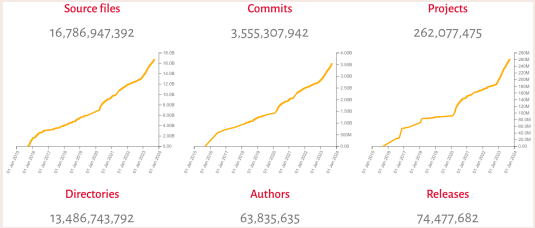
preserve and share all software source code

Research infrastructure

enable analysis of all software source code

The largest public source code archive, principled

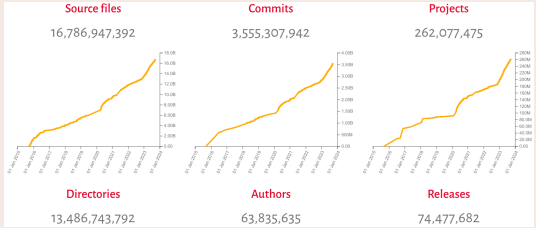
bit.ly/swhpaper



archive.softwareheritage.org

The largest public source code archive, principled

bit.ly/swhpaper



archive.softwareheritage.org

Technology

- transparency and FOSS
- replicas all the way down

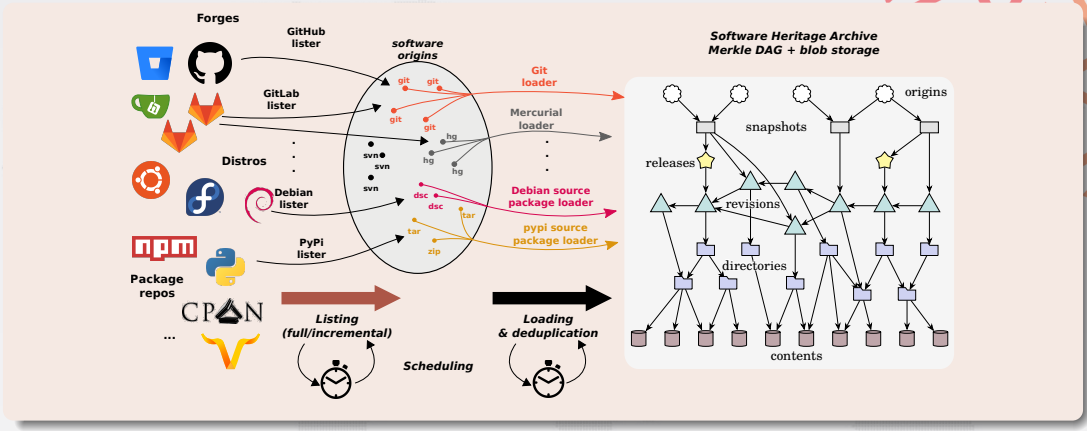
Content (billions!)

- intrinsic identifiers
- facts and provenance

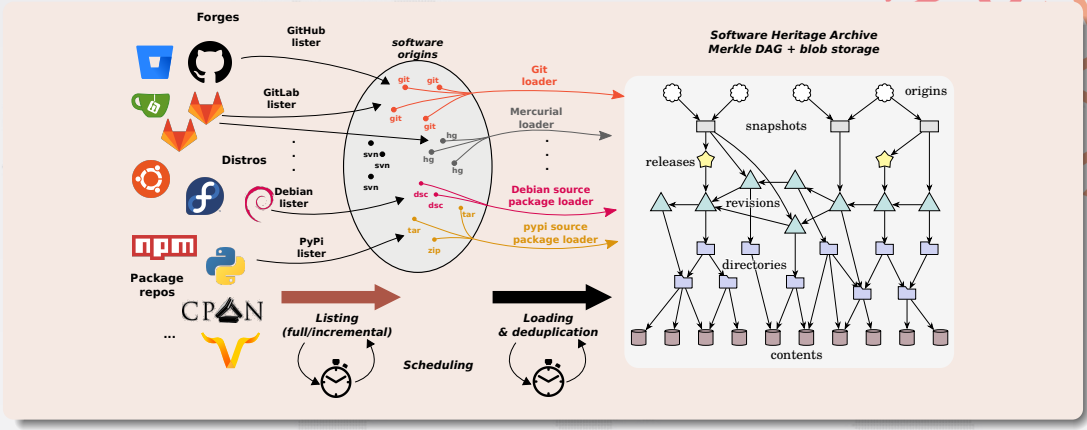
Organization

- non-profit
- multi-stakeholder

A peek under the hood: a global view on the software commons



A peek under the hood: a global view on the software commons



A global graph linking together fully deduplicated source code artifact (files, commits, directories, releases, etc.) to the places that distribute them (e.g., Git repositories), providing a unified view on the entire *Software Commons*. (Size: ~30 B nodes, ~300 B edges, ~1 PiB blobs)

Outline

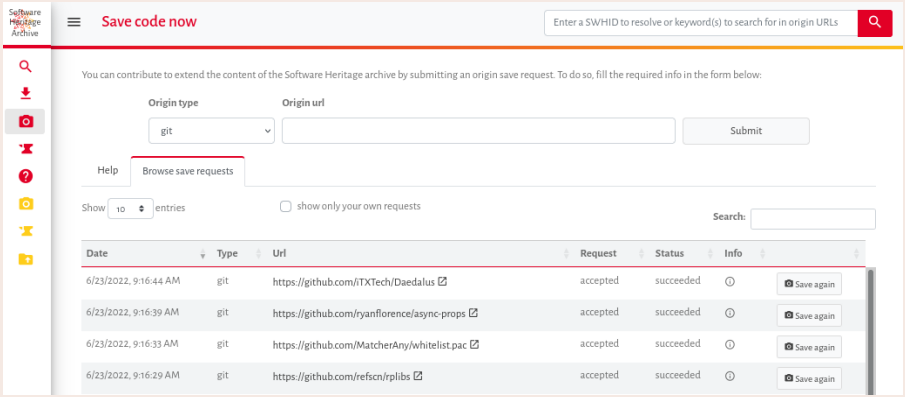
How does this relate to Guix?

- Nothing is eternal, source code (in all forms) disappears
- Hopefully, SWH keeps a copy of everything
- Guix ensures source code is archived in SWH when building ("Save Code Now")
- After source code actually disappears, falls back to SWH when rebuilding ("Software Heritage Vault")

Outline

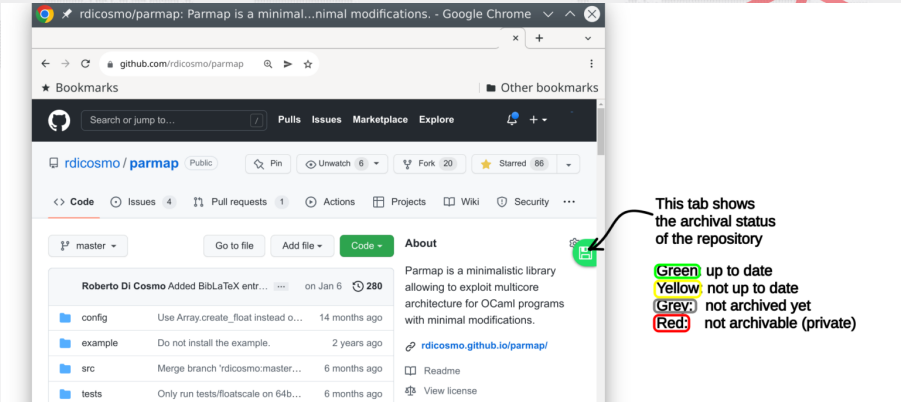
Archive a repository - Save code now

Archive a single repository



SWH Browser extension

- Directly check the availability of your repos in the archive



- HowTo and download…

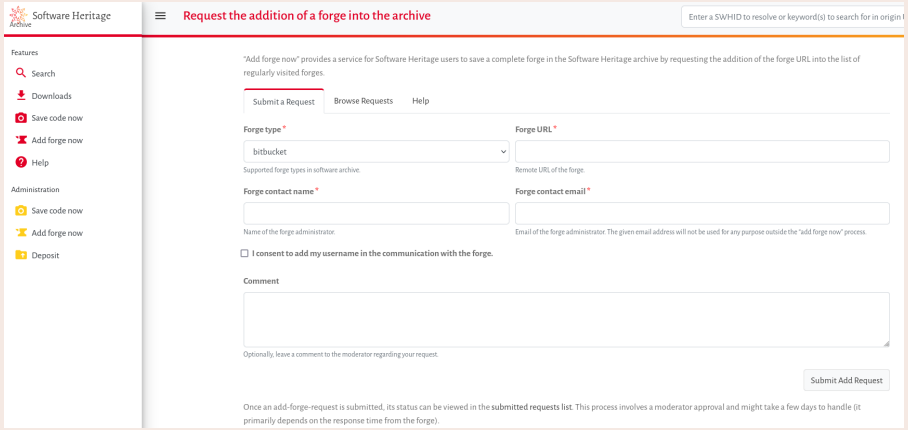
Archive a repository - Webhooks

Automate archival from your forge

- Automatically trigger archival in Software Heritage
- Triggering events:
  - Tags or branch creation
  - Release creation
  - …
- Available for:
  - Bitbucket
  - Gitea
  - GitHub
  - GitLab
  - Sourceforge
- Webhooks howto available here…

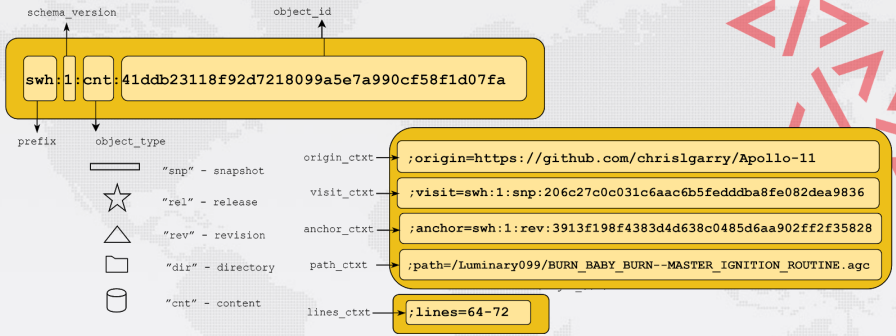
Archive a forge - Add forge now

Archive a whole forge



Outline

Software Heritage *intrinsic* Identifiers (SWHID) (full spec)



- Software Hash IDentifier - Spec v1.1

Outline

How to deposit

- Local method: deposit .zip/.tar.gz file
- SWHID method: deposit SWHID + metadata

Embedded metadata

- README file (Mardown or plain text)
- AUTHORS files (plain text)
- LICENCE file (plain text) or LICENCES directory
- codemeta.json file CodeMeta generator

CodeMeta generator

Most fields are optional. Mandatory fields will be highlighted when generating Codemeta.

The software itself

Name
My Software
the software title

Description
My Software computes ephemerides and orbit propagation. It has been developed from early '80.

Creation date
YYYY-MM-DD

First release date
YYYY-MM-DD

License(s)
from SPDX licence list

Discoverability and citation

Unique identifier
10.151.xxxxx
such as ISBNs, GTIN codes, UUIDs etc. http://schema.org/identifier

Application category
Astronomy

Keywords
ephemerides, orbit, astronomy

Funding
PRA_2018_73
grant funding software development

Funder
Università di Pisa
organization funding software development

Authors and contributors can be added below

Development community / tools

Code repository
git+https://github.com/You/RepoName.git

Continuous integration
https://travis-ci.org/You/RepoName

Issue tracker
https://github.com/You/RepoName/issues

Related links

Run-time environment

Programming Language
C#, Java, Python 3

Runtime Platform
.NET, JVM

Operating System
Android 1.6, Linux, Windows, macOS

Other software requirements
Python 3.4
https://github.com/psf/requests

Current version of the software

Version number
1.0.0

Release date
YYYY-MM-DD

Download URL
https://example.org/MySoftware.tar.gz

Release notes
Change log: this and that;
Bugfixes: that and this.

Additional Info

Reference Publication
https://doi.org/10.1000/xyz123

Development Status
see www.repostatus.org for details

Is part of
http://The.Bigger.Framework.org

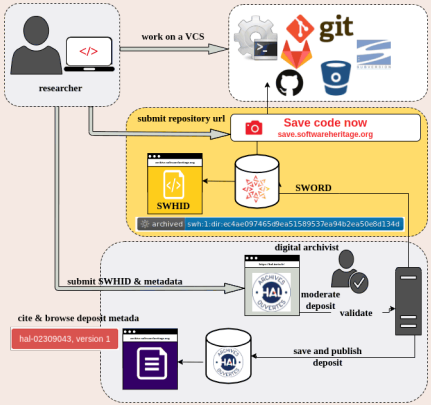
Authors

Add one | Remove last

Contributors

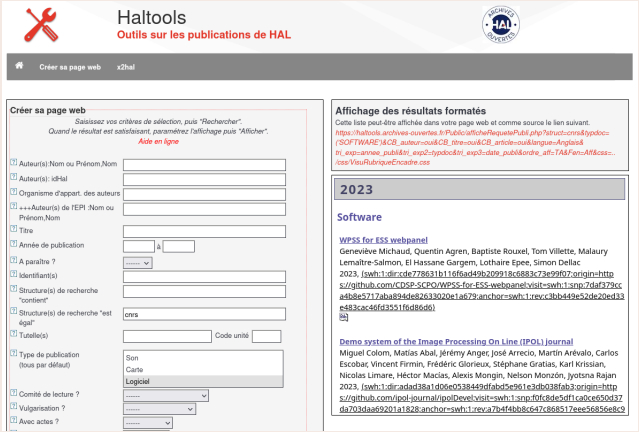
SWHID Deposit in HAL

- https://www.softwareheritage.org/2023/04/04/swhid-deposit-hal/



Search and expose software publications

- https://haltools.archives-ouvertes.fr



Outline

Questions ?

Outline

Appendix

Software Heritage

- Browse the archive
- Save and reference research software
- HOWTO archive and reference your code
- Save Code Now
- SWH browser extension
- Add Forge Now
- Webhooks for auto-archival

Misc

- CodeMeta generator
- SWHID deposit in HAL
- HAL Tools