Software Heritage, the universal archive at the crossroads of Software Forges

Benoit Chauvet

Software Heritage
Software Engineering Manager

29 June 2023
GT données GitLab 2023
Paris, France

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Introduction
2. Free Open source Software and Research
3. An Endangered Knowledge
4. Our Mission
5. Identify, cite and reference your code
6. Archive your code
7. Conclusion
8. Appendix

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

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE
We are building the universal software archive
softwareheritage.org
a project launched by Inria

Outline

1. Introduction
2. Free Open source Software and Research
3. An Endangered Knowledge
4. Our Mission
5. Identify, cite and reference your code
6. Archive your code
7. Conclusion
8. Appendix

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

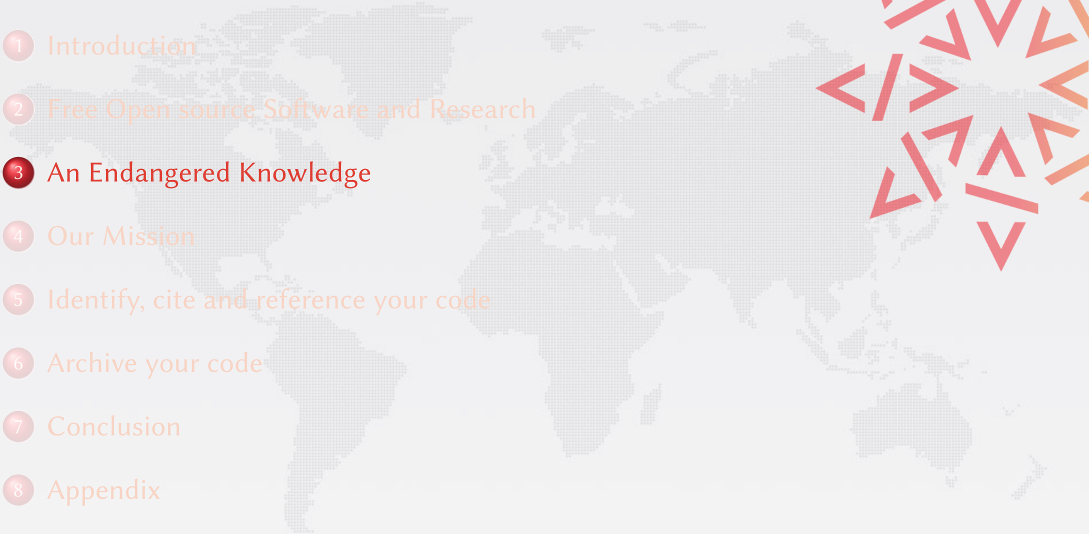
make it easy to *discover* and *reuse* them (*visibility*)

Cite/Credit

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

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

Outline



1. Introduction
2. Free Open source Software and Research
3. **An Endangered Knowledge**
4. Our Mission
5. Identify, cite and reference your code
6. Archive your code
7. Conclusion
8. Appendix

But *where* is this commons?

Sourceforge
Debian
CPAN
Gitorious
Maven
Inria
Bitbucket
GitHub
BerliOs
CTAN
Gitlab
CRAN
GoogleCode
Adullact

- 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

damage
disaster
malicious
media
obsolete
attack
aging
tear
dependencies
storage
reference
deletion
dangling
wear
corruption
encryption
format

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

damage
storage
disaster
reference
deletion
media
malicious
attack
obsolete
aging
dependencies
tear
dangling
wear
corruption
encryption
format

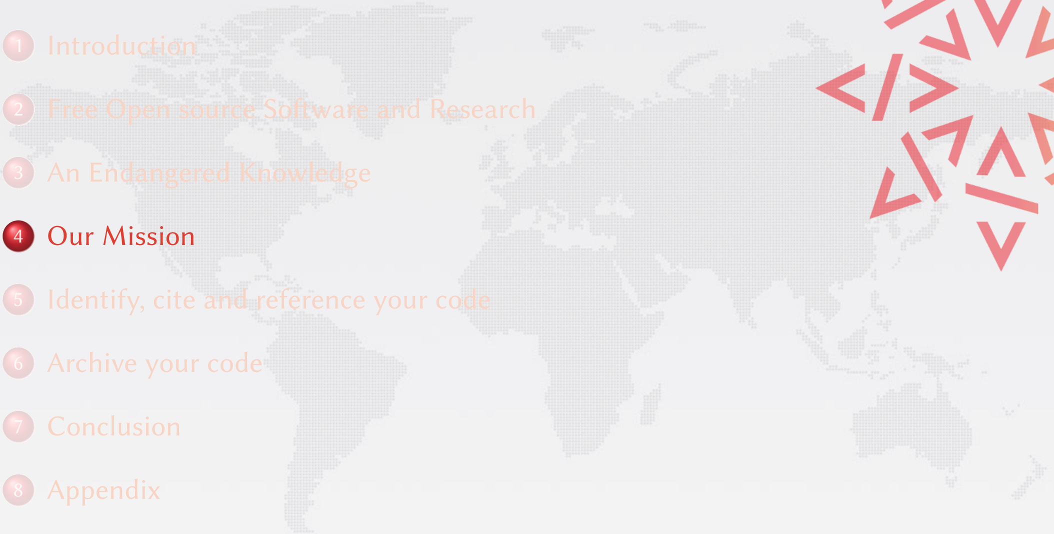
Like all digital information, FOSS is fragile

- 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



1. Introduction
2. Free Open source Software and Research
3. An Endangered Knowledge
4. **Our Mission**
5. Identify, cite and reference your code
6. Archive your code
7. Conclusion
8. Appendix

Software Heritage in a nutshell

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 in a nutshell

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

Sourceforge
Debian
CPAN
Gitorious
Maven
Inria
Bitbucket
GitHub
BerliOs
CTAN
GitLab
CRAN
GoogleCode
Adullact

find and reference all software source code

Software Heritage in a nutshell

www.softwareheritage.org

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

Debian
Sourceforge
CPAN
Gitorious
Maven
Inria
Bitbucket
GitHub
BerliOs
CTAN
GitLab
CRAN
GoogleCode
Adullact

find and reference all software source code

Universal archive

damage
storage
disaster
reference
deletion
media
malicious
obsolete
aging
attack
dependencies
tear
dangling
wear
corruption
encryption
format

preserve and share all software source code

Software Heritage in a nutshell

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

Debian
Sourceforge
CPAN
Gitorious
Maven
Inria
Bitbucket
GitHub
BerliOs
GitLab
CTAN
CRAN
GoogleCode
Adullact

find and reference all software source code

Universal archive

damage
disaster
storage
reference
deletion
malicious
media
obsolete
aging
attack
tear
dependencies
dangling
wear
corruption
encryption
format

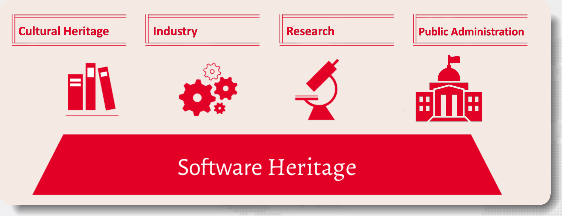
preserve and share all software source code

Research infrastructure

enable analysis of all software source code

The largest public source code archive, principled

bit.ly/swhpaper



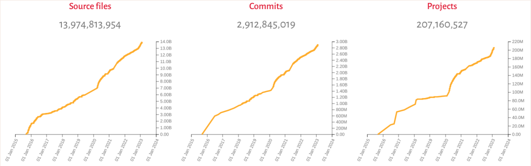
Cultural Heritage

Industry

Research

Public Administration

Software Heritage



Source files

13,974,813,954

Commits

2,912,845,019

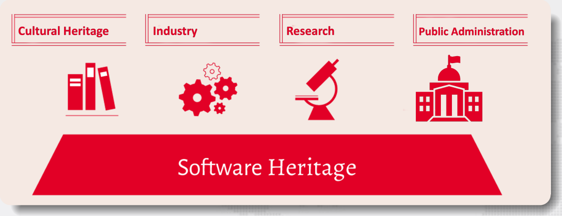
Projects

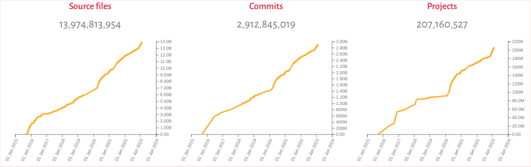
207,160,527

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

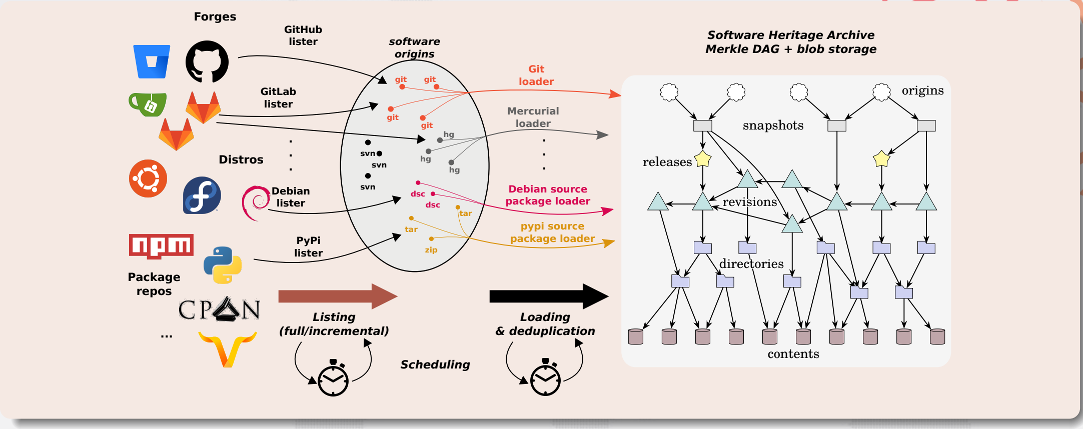
Content (billions!)

- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- multi-stakeholder

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



Forges

GitHub lister

GitLab lister

Distros

Debian lister

PyPi lister

Package repos

...

software origins

Git loader

Mercurial loader

Debian source package loader

pypi source package loader

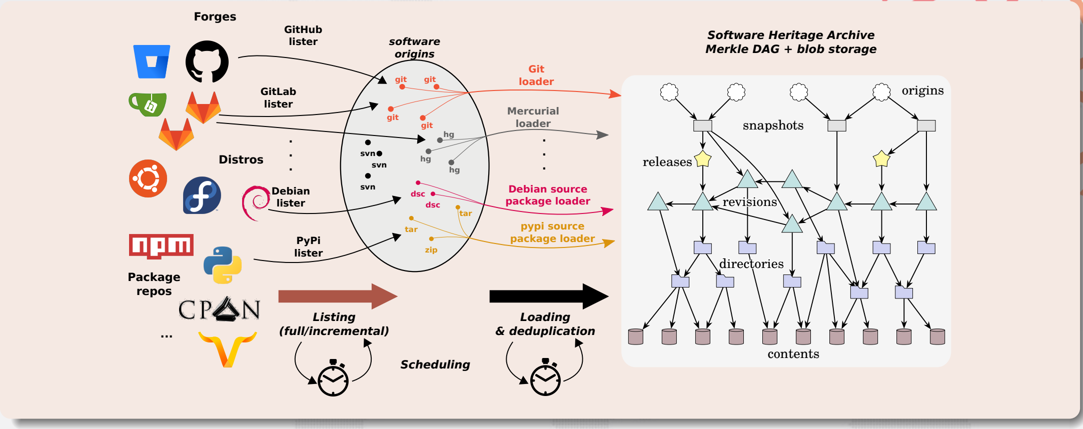
Software Heritage Archive Merkle DAG + blob storage

Listing (full/incremental)

Scheduling

Loading & deduplication

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

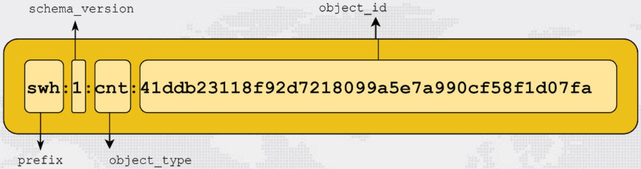


A global graph linking together fully deduplicated source code artifact (files, commits, directories, releases, etc.) to the places that distribute them (e.g., Git repositories), providing a unified view on the entire *Software Commons*. (Size: ~30 B nodes, ~300 B edges, ~1 PiB blobs)

Outline

1. Introduction
2. Free Open source Software and Research
3. An Endangered Knowledge
4. Our Mission
5. **Identify, cite and reference your code**
6. Archive your code
7. Conclusion
8. Appendix

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



schema_version

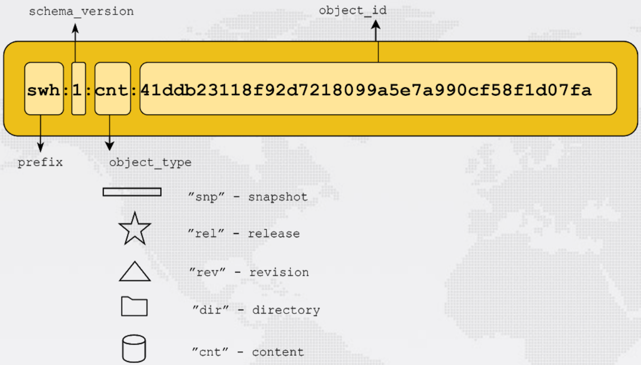
object_id

swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa

prefix

object_type

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



schema_version

object_id

swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa

prefix

object_type

”snp” - snapshot

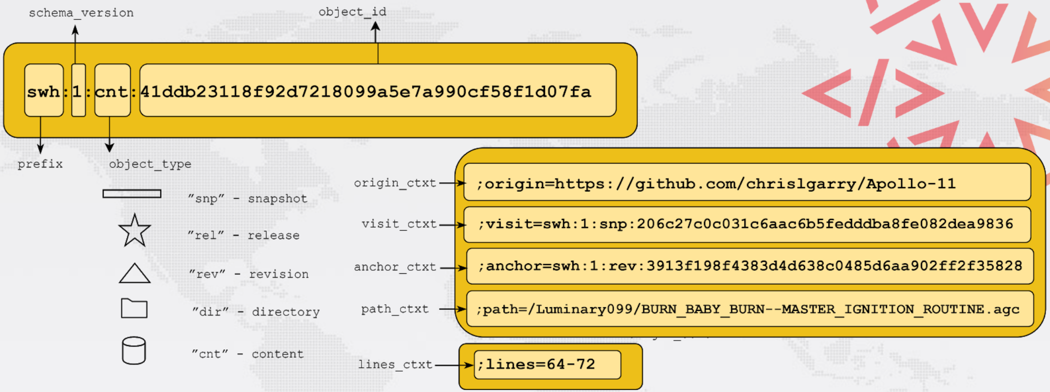
”rel” - release

”rev” - revision

”dir” - directory

”cnt” - content

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



schema_version

object_id

```
swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa
```

prefix

object_type

▭ "snp" - snapshot

☆ "rel" - release

△ "rev" - revision

🗀 "dir" - directory

⛁ "cnt" - content

origin_ctxt → ;origin=https://github.com/chrislgarry/Apollo-11

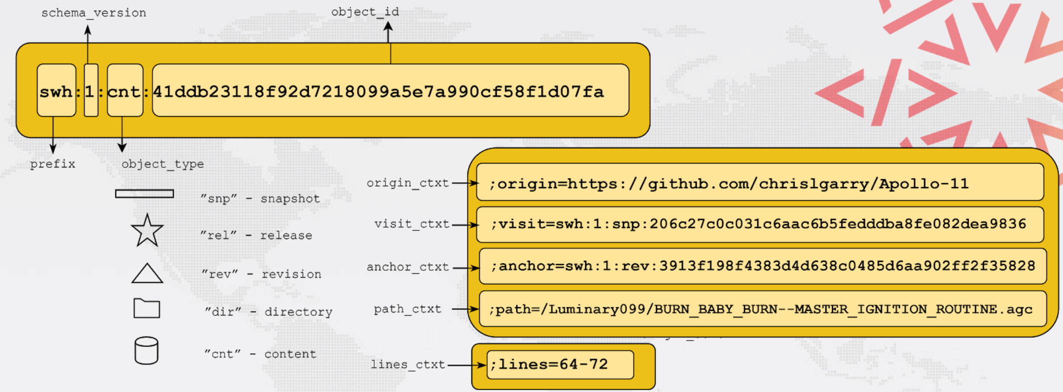
visit_ctxt → ;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836

anchor_ctxt → ;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828

path_ctxt → ;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc

lines_ctxt → ;lines=64-72

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



schema_version

object_id

swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa

prefix

object_type

"snp" - snapshot

"rel" - release

"rev" - revision

"dir" - directory

"cnt" - content

origin_ctxt → ;origin=https://github.com/chrislgarry/Apollo-11

visit_ctxt → ;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836

anchor_ctxt → ;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828

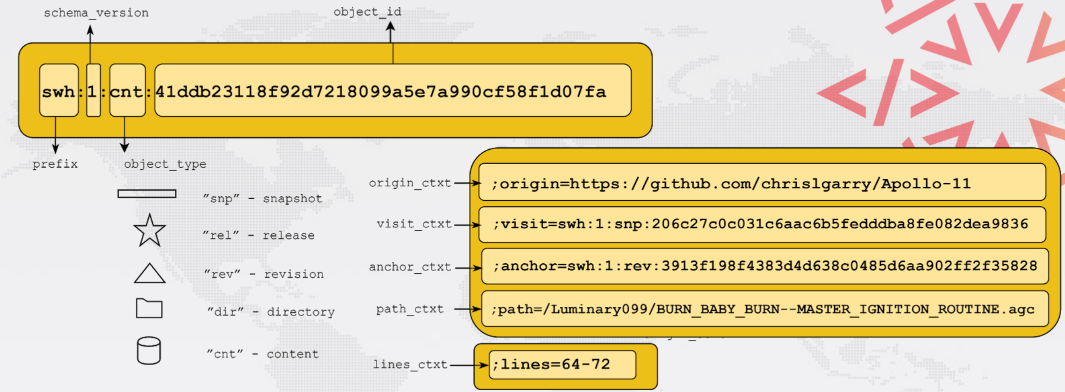
path_ctxt → ;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc

lines_ctxt → ;lines=64-72

An emerging standard

- in Linux Foundation's SPDX 2.2
- IANA registered, WikiData property P6138

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



schema_version

object_id

swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa

prefix

object_type

"snp" - snapshot

"rel" - release

"rev" - revision

"dir" - directory

"cnt" - content

origin_ctxt ;origin=https://github.com/chrislgarry/Apollo-11

visit_ctxt ;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836

anchor_ctxt ;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828

path_ctxt ;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc

lines_ctxt ;lines=64-72

An emerging standard

- in Linux Foundation's SPDX 2.2
- IANA registered, WikiData property P6138

Examples:

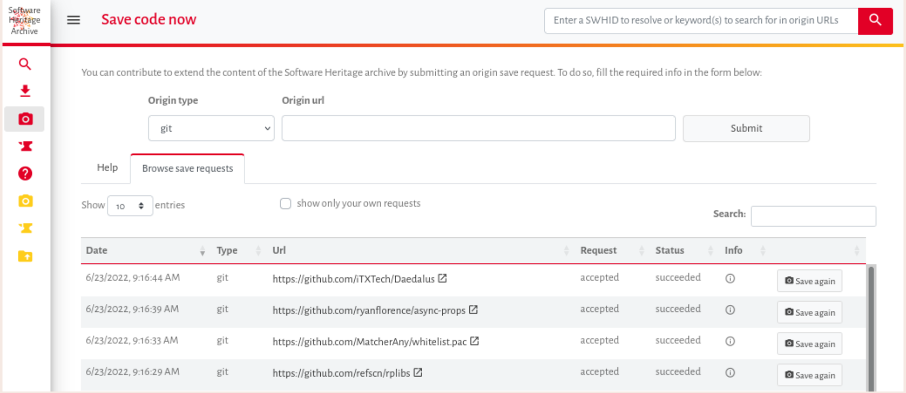
- Apollo 11 AGC excerpt
- Quake III rsqrt

Outline

1. Introduction
2. Free Open source Software and Research
3. An Endangered Knowledge
4. Our Mission
5. Identify, cite and reference your code
6. Archive your code
7. Conclusion
8. Appendix

Archive a repository - Save code now

Archive a single repository



Software Heritage Archive
Save code now
Enter a SWHID to resolve or keyword(s) to search for in origin URLs

You can contribute to extend the content of the Software Heritage archive by submitting an origin save request. To do so, fill the required info in the form below:

Origin type: git

Origin url:

Submit
Help
Browse save requests

Show 10 entries

□ show only your own requests

Search:

Date	Type	Url	Request	Status	Info	
6/23/2022, 9:16:44 AM	git	https://github.com/iTXTech/Daedalus	accepted	succeeded	ⓘ	Save again
6/23/2022, 9:16:39 AM	git	https://github.com/ryanflorence/async-props	accepted	succeeded	ⓘ	Save again
6/23/2022, 9:16:33 AM	git	https://github.com/MatcherAny/whitelist.pac	accepted	succeeded	ⓘ	Save again
6/23/2022, 9:16:29 AM	git	https://github.com/refscn/rplibs	accepted	succeeded	ⓘ	Save again

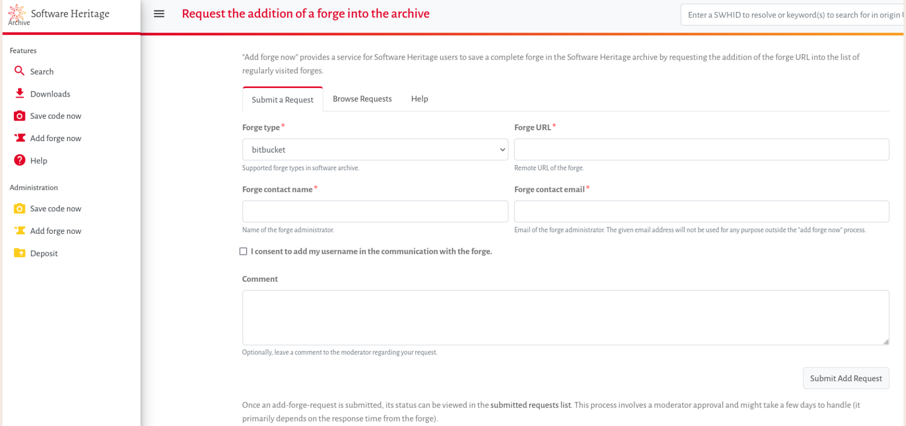
Archive a repository - Webhooks

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



Request the addition of a forge into the archive

"Add forge now" provides a service for Software Heritage users to save a complete forge in the Software Heritage archive by requesting the addition of the forge URL into the list of regularly visited forges.

Forge type * : bitbucket
Supported forge types in software archive.

Forge URL * :
Remote URL of the forge.

Forge contact name * :
Name of the forge administrator.

Forge contact email * :
Email of the forge administrator. The given email address will not be used for any purpose outside the "add forge now" process.

☐ I consent to add my username in the communication with the forge.

Comment:
Optionally, leave a comment to the moderator regarding your request.

Once an add-forge-request is submitted, its status can be viewed in the **submitted requests list**. This process involves a moderator approval and might take a few days to handle (it primarily depends on the response time from the forge).

HAL Deposit

HAL

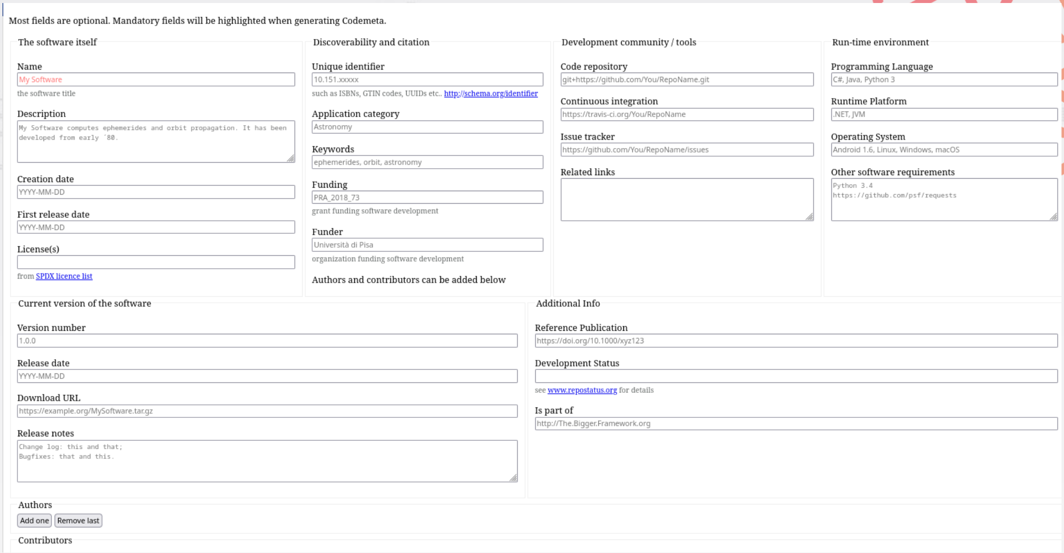
How to deposit

- 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

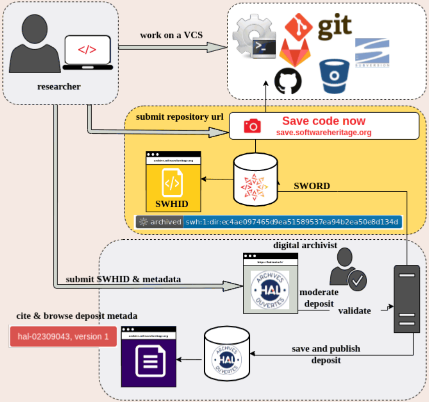
Authors

Contributors

HAL deposit with SWHID method

SWHID Deposit in HAL

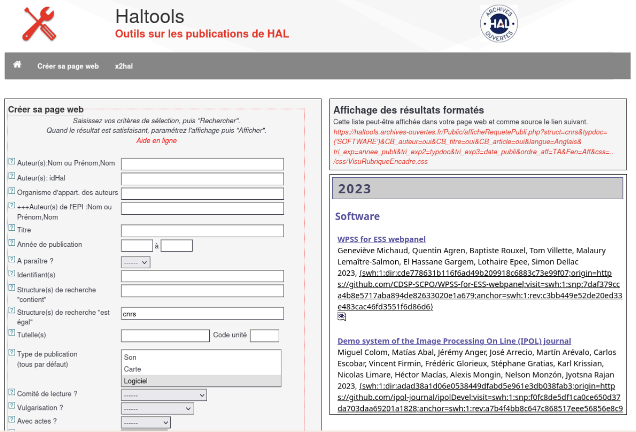
- https://www.softwareheritage.org/2023/04/04/swhid-deposit-hal/



HAL tools

Search and expose software publications

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



Créer sa page web

Saisissez vos critères de sélection, puis "Rechercher".
Quand le résultat est satisfaisant, paramétrez l'affichage puis "Afficher".
Aide en ligne

Auteur(s):Nom ou Prénom,Nom	
Auteur(s): idHal	
Organisme d'appart. des auteurs	
+++Auteur(s) de l'EPI :Nom ou Prénom,Nom	
Titre	
Année de publication	à
A paraître ?	------
Identifiant(s)	
Structure(s) de recherche "contient"	
Structure(s) de recherche "est égal"	cnrs
Tutelle(s)	Code unité
Type de publication (tous par défaut)	Son Carte Logiciel
Comité de lecture ?	------
Vulgarisation ?	------
Avec actes ?	------

Affichage des résultats formatés

Cette liste peut-être affichée dans votre page web et comme source le lien suivant.
https://haltools.archives-ouvertes.fr/Public/afficheRequetePubli.php?struct=cnrs&typdoc=('SOFTWARE')&CB_auteur=oui&CB_titre=oui&CB_article=oui&langue=Anglais&tri_exp=annee_publi&tri_exp2=typdoc&tri_exp3=date_publi&ordre_aff=TA&Fen=Aff&css=../css/VisuRubriqueEncadre.css

2023

Software

WPSS for ESS webpanel
Geneviève Michaud, Quentin Agren, Baptiste Rouxel, Tom Villette, Malaury Lemaître-Salmon, El Hassane Gargem, Lothaire Epee, Simon Dellac
2023, ⟨swh:1:dir:cde778631b116f6ad49b209918c6883c73e99f07;origin=https://github.com/CDSP-SCPO/WPSS-for-ESS-webpanel;visit=swh:1:snp:7daf379cca4b8e5717aba894de82633020e1a679;anchor=swh:1:rev:c3bb449e52de20ed33e483cac46fd3551f6d86d6⟩
Bib

Demo system of the Image Processing On Line (IPOL) journal
Miguel Colom, Matías Abal, Jérémy Anger, José Arrecio, Martín Arévalo, Carlos Escobar, Vincent Firmin, Frédéric Glorieux, Stéphane Gratias, Karl Krissian, Nicolas Limare, Héctor Macías, Alexis Mongin, Nelson Monzón, Jyotsna Rajan
2023, ⟨swh:1:dir:adad38a1d06e0538449dfabd5e961e3db038fab3;origin=https://github.com/ipol-journal/ipolDevel;visit=swh:1:snp:f0fc8de5df1ca0ce650d37da703daa69201a1828;anchor=swh:1:rev:a7b4f4bb8c647c868517eee56856e8c9

Outline

1. Introduction
2. Free Open source Software and Research
3. An Endangered Knowledge
4. Our Mission
5. Identify, cite and reference your code
6. Archive your code
7. Conclusion
8. Appendix

Questions ?

Outline

1. Introduction
2. Free Open source Software and Research
3. An Endangered Knowledge
4. Our Mission
5. Identify, cite and reference your code
6. Archive your code
7. Conclusion
8. Appendix

Appendix

Software Heritage

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

Misc

- CodeMeta generator
- SWHID deposit in HAL
- HAL Tools