

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

- 1 Introduction
- 2 Free Open source Software and Research
- 3 An Endangered Knowledge
- 4 Our Mission
- 5 Collaboration Guix / SWH
- 6 Archive your code
- 7 Identify, cite and reference your code
- 8 HAL deposit
- 9 Conclusion
- 10 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



- 1 Introduction
- 2 Free Open source Software and Research
- 3 An Endangered Knowledge
- 4 Our Mission
- 5 Collaboration Guix / SWH
- 6 Archive your code
- 7 Identify, cite and reference your code
- 8 HAL deposit
- 9 Conclusion
- 10 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!*)

- 1 Introduction
- 2 Free Open source Software and Research
- 3 An Endangered Knowledge**
- 4 Our Mission
- 5 Collaboration Guix / SWH
- 6 Archive your code
- 7 Identify, cite and reference your code
- 8 HAL deposit
- 9 Conclusion
- 10 Appendix



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



A word cloud of terms related to software fragility, including: damage, disaster, malicious, attack, obsolete, dependencies, deletion, reference, storage, dangling, wear, corruption, encryption, and format. The words are arranged in a circular pattern with varying sizes and colors.

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



A word cloud centered on a world map background. The most prominent words are 'damage', 'disaster', 'malicious', 'attack', 'obsolete', 'deletion', and 'format'. Other smaller words include 'reference', 'storage', 'dependencies', 'dangling', 'wear', 'corruption', 'encryption', 'tear', 'aging', and 'media'. The words are in various colors and orientations, with 'format' being the largest and most vertical.

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?

- 1 Introduction
- 2 Free Open source Software and Research
- 3 An Endangered Knowledge
- 4 Our Mission**
- 5 Collaboration Guix / SWH
- 6 Archive your code
- 7 Identify, cite and reference your code
- 8 HAL deposit
- 9 Conclusion
- 10 Appendix





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 Software Heritage logo is a red trapezoid with the text "Software Heritage" in white. Above it are four categories, each with an icon: Cultural Heritage (books), Industry (gears), Research (microscope), and Public Administration (government building).

Category	Icon
Cultural Heritage	Books
Industry	Gears
Research	Microscope
Public Administration	Government Building



archive.softwareheritage.org



archive.softwareheritage.org

Technology

- transparency and FOSS
- replicas all the way down

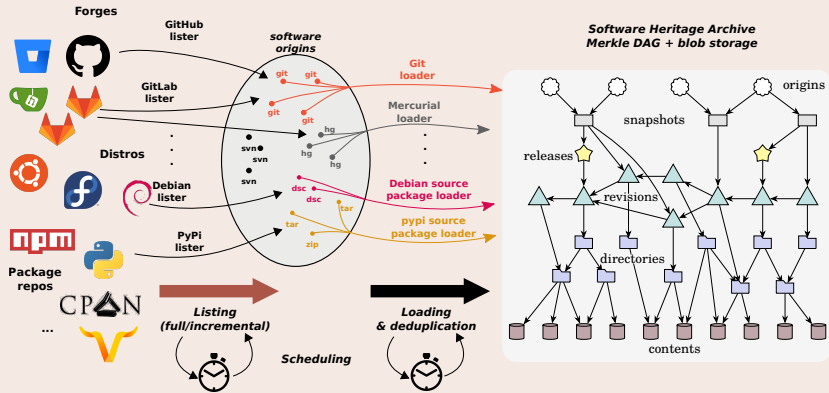
Content (billions!)

- intrinsic identifiers
- facts and provenance

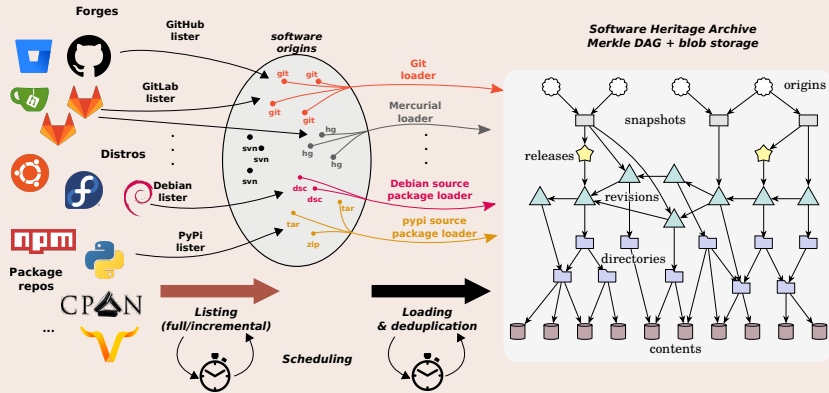
Organization

- 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)

- 1 Introduction
- 2 Free Open source Software and Research
- 3 An Endangered Knowledge
- 4 Our Mission
- 5 Collaboration Guix / SWH**
- 6 Archive your code
- 7 Identify, cite and reference your code
- 8 HAL deposit
- 9 Conclusion
- 10 Appendix



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")



- 1 Introduction
- 2 Free Open source Software and Research
- 3 An Endangered Knowledge
- 4 Our Mission
- 5 Collaboration Guix / SWH
- 6 Archive your code**
- 7 Identify, cite and reference your code
- 8 HAL deposit
- 9 Conclusion
- 10 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:

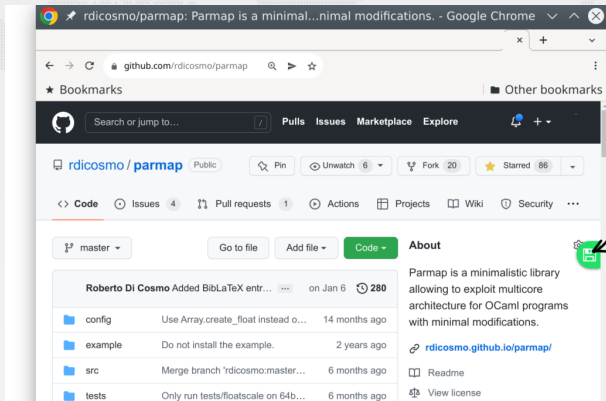
Origin type: Origin url:

Help: [Browse save requests](#)

Show entries show only your own requests Search:

Date	Type	Url	Request	Status	Info
6/23/2022, 9:16:44 AM	git	https://github.com/iTXXTech/Daedalus	accepted	succeeded	<input type="button" value="Save again"/>
6/23/2022, 9:16:39 AM	git	https://github.com/ryanflorence/async-props	accepted	succeeded	<input type="button" value="Save again"/>
6/23/2022, 9:16:33 AM	git	https://github.com/MatcherAny/whitelist.pac	accepted	succeeded	<input type="button" value="Save again"/>
6/23/2022, 9:16:29 AM	git	https://github.com/refscn/rplibs	accepted	succeeded	<input type="button" value="Save again"/>

- Directly check the availability of your repos in the archive



This tab shows
the archival status
of the repository

Green up to date
Yellow not up to date
Grey not archived yet
Red not archivable (private)

- [HowTo and download...](#)

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 whole forge

Software Heritage

Request the addition of a forge into the archive

Enter a SWHID to resolve or keyword(s) to search for in origin

"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.

Submit a Request Browse Requests Help

Forge type * bitbucket

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.

Submit Add 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).

- 1 Introduction
- 2 Free Open source Software and Research
- 3 An Endangered Knowledge
- 4 Our Mission
- 5 Collaboration Guix / SWH
- 6 Archive your code
- 7 Identify, cite and reference your code**
- 8 HAL deposit
- 9 Conclusion
- 10 Appendix





- Software Hash Identifier - [Spec v1.1](#)

- 1 Introduction
- 2 Free Open source Software and Research
- 3 An Endangered Knowledge
- 4 Our Mission
- 5 Collaboration Guix / SWH
- 6 Archive your code
- 7 Identify, cite and reference your code
- 8 HAL deposit**
- 9 Conclusion
- 10 Appendix





How to deposit

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

Embedded metadata

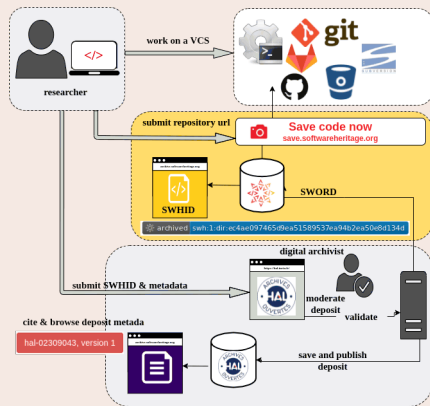
- README file (Markdown or plain text)
- AUTHORS files (plain text)
- LICENCE file (plain text) or LICENCES directory
- `codemeta.json` file [CodeMeta generator](#)

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

The software itself Name <input type="text" value="My Software"/> the software title Description <input type="text" value="My Software computes ephemerides and orbit propagation. It has been developed from early '88."/> Creation date <input type="text" value="YYYY-MM-DD"/> First release date <input type="text" value="YYYY-MM-DD"/> License(s) <input type="text"/> from SPDX licence list	Discoverability and citation Unique identifier <input type="text" value="10.151.xxxxx"/> such as ISBNs, GTIN codes, UUIDs etc. http://schema.org/identifier Application category <input type="text" value="Astronomy"/> Keywords <input type="text" value="ephemerides, orbit, astronomy"/> Funding <input type="text" value="PRA_2018_73"/> grant funding software development Funder <input type="text" value="Università di Pisa"/> organization funding software development Authors and contributors can be added below	Development community / tools Code repository <input type="text" value="git+https://github.com/You/RepoName.git"/> Continuous integration <input type="text" value="https://travis-ci.org/You/RepoName"/> Issue tracker <input type="text" value="https://github.com/You/RepoName/issues"/> Related links <input type="text"/>	Run-time environment Programming Language <input type="text" value="C#, Java, Python 3"/> Runtime Platform <input type="text" value=".NET, JVM"/> Operating System <input type="text" value="Android 1.6, Linux, Windows, macOS"/> Other software requirements <input type="text" value="Python 3.4
https://github.com/psf/requests"/>
Current version of the software Version number <input type="text" value="1.0.0"/> Release date <input type="text" value="YYYY-MM-DD"/> Download URL <input type="text" value="https://example.org/MySoftware.tar.gz"/> Release notes <input type="text" value="Change log: this and that;
Bugfixes: that and this."/> Authors <input type="button" value="Add one"/> <input type="button" value="Remove last"/> Contributors	Additional Info Reference Publication <input type="text" value="https://doi.org/10.1000/xyz123"/> Development Status <input type="text"/> see www.repostatus.org for details Is part of <input type="text" value="http://The.Bigger.Framework.org"/>		


SWHID Deposit in HAL

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




Search and expose software publications

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



Haltools

Outils sur les publications de HAL



Créer sa page web
x2hal

Créer sa page web

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

Auteurs(s) Nom ou Prénom Nom

Auteurs(s) Id-Hal

Organisme d'appart. des auteurs

+++Auteurs(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"

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/PublicAfficheRequetePubli.php?struc=cns&typdoc=\(SOFTWARE\)&CB_auteur=oui&CB_titre=oui&CB_article=oui&langue=Anglais&tr_exp=annee_public&tr_exp2=typdoc&tr_exp3=date_public&ordre_aff=TA&Fan=Aff&css=css-VisuRubriqueEncadre.css](https://haltools.archives-ouvertes.fr/PublicAfficheRequetePubli.php?struc=cns&typdoc=(SOFTWARE)&CB_auteur=oui&CB_titre=oui&CB_article=oui&langue=Anglais&tr_exp=annee_public&tr_exp2=typdoc&tr_exp3=date_public&ordre_aff=TA&Fan=Aff&css=css-VisuRubriqueEncadre.css)

2023

Software

[WPSS for ESS webpanel](#)
 Geneviève Michaud, Quentin Agren, Baptiste Rouxel, Tom Villette, Malaury Lemaitre-Salmori, El Hassane Gargem, Lothaire Epee, Simon Dellac
 2023, [swh:1:dir:cde778631b116f6ad49b209918c6883c73e99f07:origin=http://github.com/CDSP-SCPO/WPSS-for-ESS-webpanel/visit-swh:1:snp:7daf379cc448e5717aba894de82633020e1a679:anchor=swh:1:rev:c3bb449e52de20ed33e483cac46fd3551fd86d6]

[Demo system of the Image Processing On Line \(IPOL\) Journal](#)
 Miguel Colom, Matias 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=http://github.com/ipol-journal/ipolDevel/visit-swh:1:snp:f0fc8de5df1cabce650d37da703daa69201a1828:anchor=swh:1:rev:a7b4f4bb8c647c868517ee56856e8c9]

- 1 Introduction
- 2 Free Open source Software and Research
- 3 An Endangered Knowledge
- 4 Our Mission
- 5 Collaboration Guix / SWH
- 6 Archive your code
- 7 Identify, cite and reference your code
- 8 HAL deposit
- 9 Conclusion**
- 10 Appendix





Questions ?

- 1 Introduction
- 2 Free Open source Software and Research
- 3 An Endangered Knowledge
- 4 Our Mission
- 5 Collaboration Guix / SWH
- 6 Archive your code
- 7 Identify, cite and reference your code
- 8 HAL deposit
- 9 Conclusion
- 10 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](#)