Archive, identify, describe and cite software source code with HAL and with Software Heritage

Morane Gruenpeter, Inria
Alexis Lebis, IMT Lille Douai

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

- ★ **Introduction**
- ★ Preserving source code
- ★ Software Heritage the universal source code archive
- ★ Research Software: a first class research output
- ★ Good practices for software curation
- ★ Conclusion

Morane Gruenpeter

Software engineer and metadata specialist

Timeline:

- 2008-2011 B.A in **Musique** (Harpist)
- 2012-2015 Licence (B.SC) in **Computer Science** @CNAM
- 2015-2017 Master (M.SC) in **Software Engineering** (R&D) @UPMC
- 2017 Internship **Software Heritage** (SWH)
- 2018-2019 European project EU2020 **CROSSMINER**(on SWH team)
- 2020-2022 European project **FAIRsFAIR** (on SWH team, @Inria)
- 2022-2025 European project **FAIRCORE4EOSC** (on SWH team, @Inria)

Working groups for Open Science and digital preservation

- the Research Data Alliance's **Software Source Code** Interest Group (SSC IG),
- the FORCE11's **Software Citation** Implementation Working Group (SCI WG),
- the joint RDA, ReSA & FORCE11 **FAIR for Research Software** Working Group (FAIR4RS WG)
- WikiData for **Digital Preservation** initiative (WikiDigi).

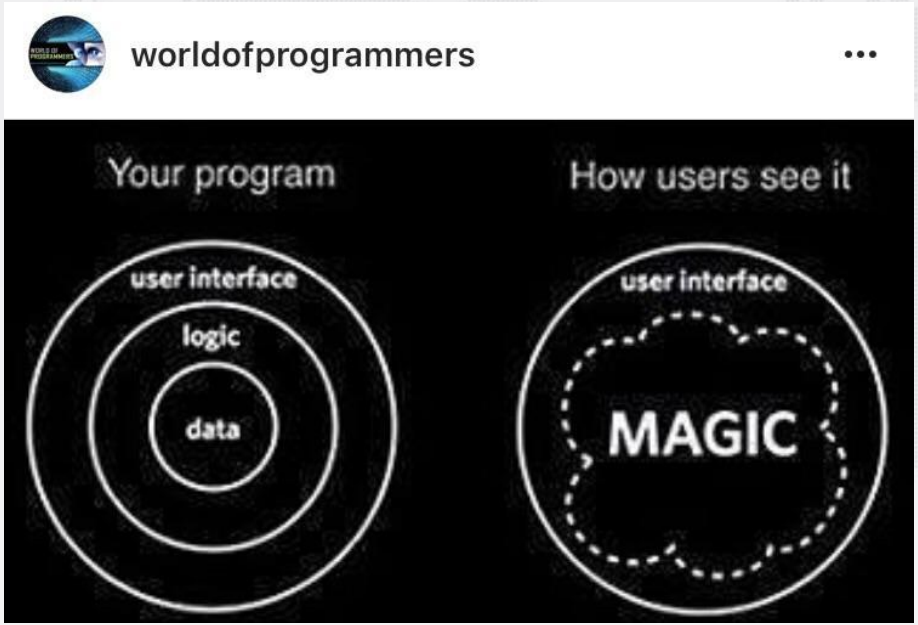
Alexis Lebis

AI LECTURER AT CERI SN OF INSTITUT MINES TÉLÉCOM LILLE DOUAI

Alexis Lebis a lecturer in artificial intelligence at the Center for Digital Systems (CERI SN) of Institut Mines Télécom Lille Douai, in France, since 2020. Before that, he was a post-doctorate for the APACHES (FIPE18-007-VERMEULEN) project.

- Decision Making (DM) and Knowledge Engineering (KE), especially applied to Technology Enhanced Learning (T.E.L)
- new decision-making tools for complex pedagogical situations,
- modeling complex (and combinatorial) TEL problems
- reproducible science, and how we can foster it in T.EL.
- Software Heritage ambassador from 2021.

What is software?



Software as a concept

- **project** or entity
- the **community** around the project
- the software **idea** / algorithms / solutions

Software artifacts

- Executables
- Source code

Ceci n'est pas une pipe.

What about **software source code** ?

Program (excerpt of binary)

```
4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
4004f9: 90
4004fa: 5d
4004fb: c3
```

Program (source code)

```
/* Hello World program */

#include<stdio.h>

void main()
{
      printf("Hello World");
}
```

Hello World

The Knowledge is in the Source Code

“Programs must be written for people to read, and only incidentally for machines to execute.”

Harold Abelson, 1985

Structure and Interpretation of Computer Programs (1st ed.),

“Source code provides a view into the mind of the designer.”

Len Shustek, 2006

Computer History Museum

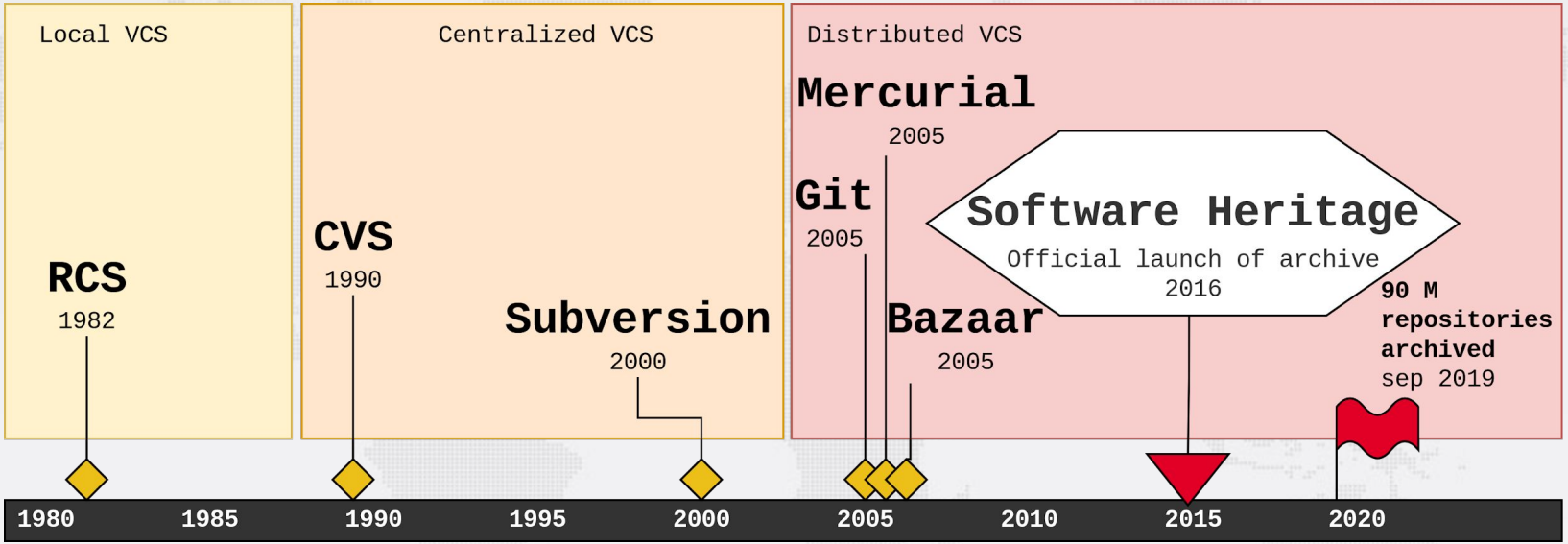
“The source code for a work means the preferred form of the work for making modifications to it."

GPL Licence

Outline

- ★ Introduction
- ★ **Preserving source code**
- ★ Software Heritage the universal source code archive
- ★ Research Software: a first class research output
- ★ Good practices for software curation
- ★ Conclusion

Version control system (VCS) history



- records changes made to a (set of) source code file (s)
- allows to operate on versions: diff/merge/fork/recover etc.
- essential tool for software development

Why source code is fragile?

Source code can be destroyed

Google Kills Off Google Code

Natasha Lomas @riptari / 10:58 AM GMT+1 • March 13, 2015

Source: TechCrunch

1.4 million projects

Posted: Thursday, March 12, 2015

When we started the Google Code project hosting service in 2006, the world of project hosting was limited. We were worried about reliability and stagnation, so we took action by giving the open source community another option to choose from. Since then, we've seen a wide variety of better project hosting services such as GitHub and Bitbucket bloom. Many projects moved away from Google Code to those other systems. To meet developers where they are, we ourselves migrated nearly a thousand of our own open source projects from Google Code to GitHub.

As developers migrated away from Google Code, a growing share of the remaining projects were spam or abuse. Lately, the administrative load has consisted almost exclusively of abuse management. After profiling non-abusive activity on Google Code, it has become clear to us that the service simply isn't needed anymore.

Beginning today, we have disabled new project creation on Google Code. We will be shutting down the service about 10 months from now on January 25th, 2016. Below, we provide links to migration tools designed to help you move your projects off of Google Code. We will also make ourselves available over the next three months to those projects that need help migrating from Google Code to other hosts.

- March 12, 2015 - New project creation disabled.
- August 24, 2015 - The site goes read-only. You can still checkout/view project source, issues, and wikis.
- January 25, 2016 - The project hosting service is closed. You will be able to download a tarball of project source, issues, and wikis. These tarballs will be available throughout the rest of 2016.

Google will continue to provide Git and Gerrit hosting for certain projects like Android and Chrome. We will also continue maintaining our mirrors of projects like Eclipse, kernel.org and others.

In science, reproducibility requires long-term access to source code

Source: BitBucket blog

250,000 repos

GitHub now belongs to Microsoft

Microsoft to acquire GitHub for $7.5 billion

June 4, 2018 | Microsoft News Center

Acquisition will empower developers, accelerate GitHub's growth and advance Microsoft services with new audiences

Source: Microsoft Blog

REUTERS

World Business Markets Breakingviews Video More

TECHNOLOGY, MEDIA & TELECOM - INNOVATION JUNE 5, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

GitLab gains developers after Microsoft buys rival GitHub

By Vibhuti Sharma, Supantha Mukherjee 3 MIN READ

Source: Reuters

Hosting your open-source project

- on a **free**, **publicly available** platform is fine.
- But you have to prepare for the platform shutdown (you need a plan B).

Source: Geralt, Pixabay

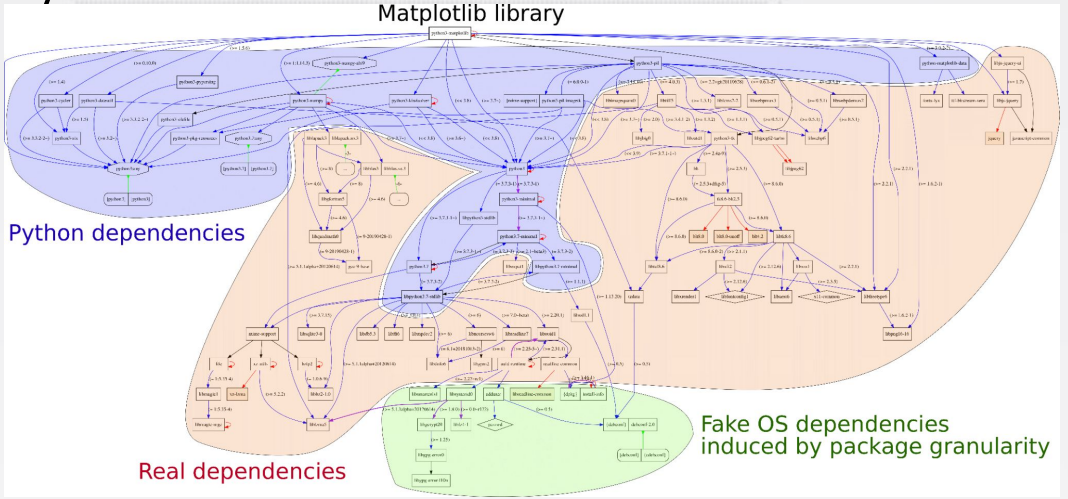
The development history is key

Software evolves over time

- projects may last decades
- the development history is key to its understanding

Complexity

- millions of lines of code
- large web of dependencies
- easy to break, difficult to maintain
- sophisticated developer communities



P. Alliez, R. Di Cosmo, B. Guedj, A. Girault, M.-S. Hacid, et al.. Attributing and Referencing (Research) Software: Best Practices and Outlook from Inria. Computing in Science and Engineering, Institute of Electrical and Electronics Engineers, 2019, pp.1-14. 〈10.1109/MCSE.2019.2949413〉. (hal-02135891)

Software identification

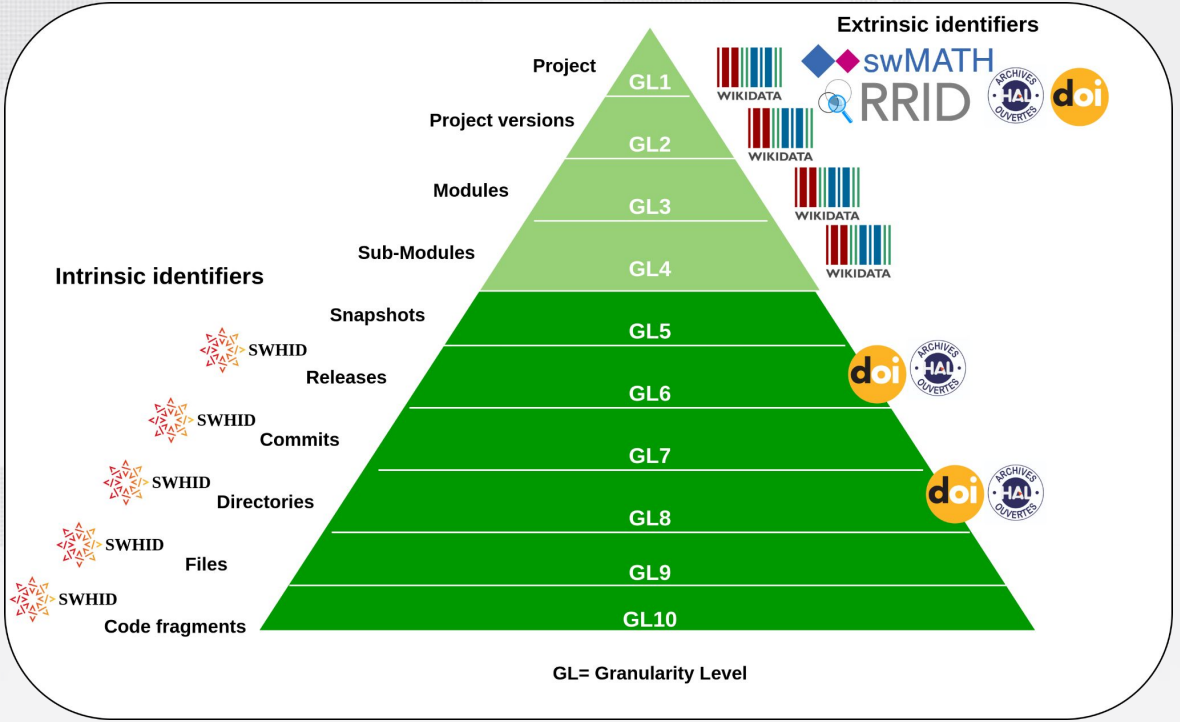
Software concept / project / collection

Description in registry, a homepage or any other form of metadata record

- Project versions (for example Python2 and Python3)
- Modules
- Sub-modules

Software artifact

- Executable (download link)
- Software source code
  - Dynamic artifact - current development code
  - Archived copy
    - Snapshot (all branches, all dev history)
    - Release / Package
    - Commit- a specific point in development history
    - Directory
    - File
    - Algorithm



Research Data Alliance/FORCE11 Software Source Code Identification WG et al. (2020). Use cases and identifier schemes for persistent software source code identification (V1.1). *Research Data Alliance*. https://doi.org/10.15497/RDA00053

Outline

- ★ Introduction
- ★ Preserving source code
- ★ **Software Heritage the universal source code archive**
- ★ Research Software: a first class research output
- ★ Good practices for software curation
- ★ Conclusion

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

- Non-profit organization
- Launched in 2016 by INRIA (Roberto Di Cosmo & Stefano Zacchiroli)
- Archives publicly available code permanently and for free.

Software Heritage in a nutshell

Reference catalog

find and reference all software source code

Universal archive

preserve all software source code

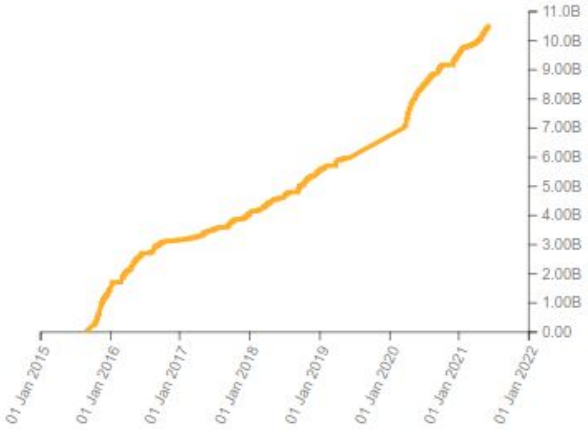
Research infrastructure

enable analysis of all software source code

In numbers https://archive.softwareheritage.org/

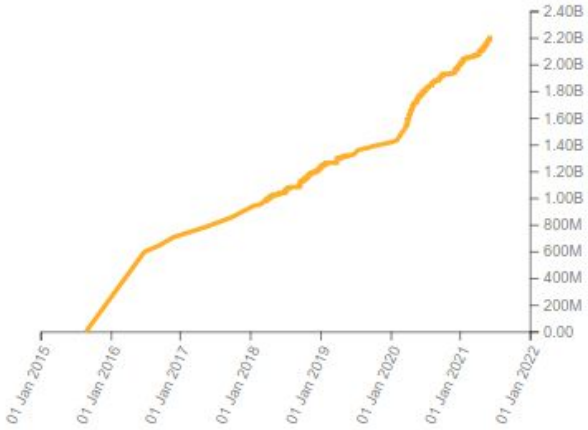
Source files

10545239307



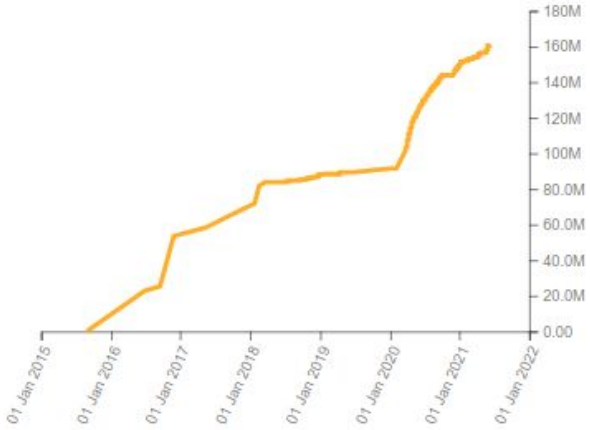
Commits

2215028000



Projects

161168065



Directories

8776445689

Authors

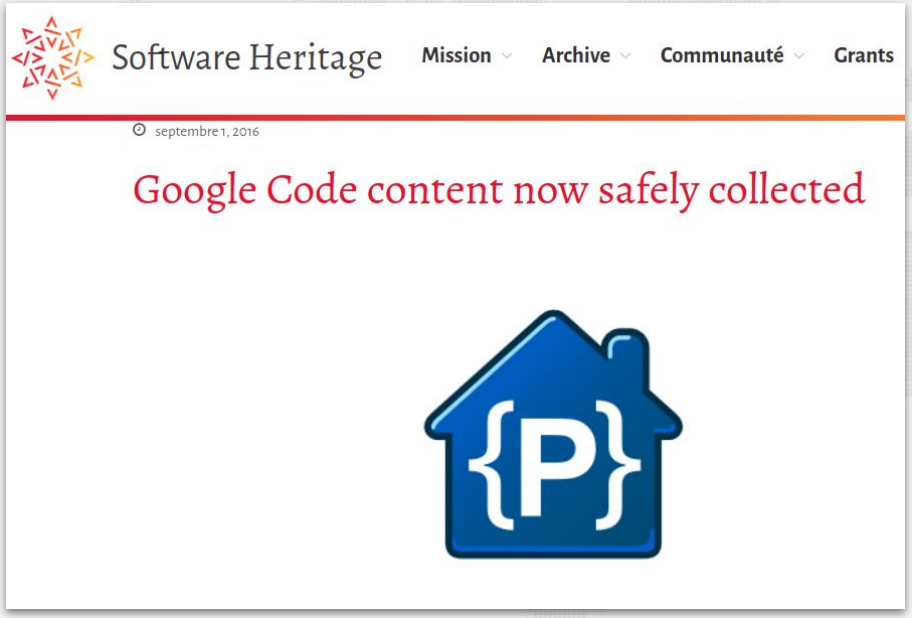
43822776

Releases

17972096

Source: Software Heritage (June 2021)

Rescuing software



Source: Software Heritage

Source: Software Heritage

An international, non profit initiative

Diamond sponsor

cea

Platinum sponsors

cnrs HUAWEI intel

Microsoft MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION SOCIETE GENERALE

Gold sponsors

openinventionnetwork Université de Paris SORBONNE UNIVERSITÉ

Silver sponsors

AdaCore CAST Software Intelligence for Digital Leaders Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE Google

GitHub UNIVERSITÀ DI PISA vmware

Bronze sponsors

Data Archiving and Networked Services DANS FOSSID

How to archive the world's source code?

Inside of Library of Alexandria-artistic rendition (Image: 21h007)

Archiving software

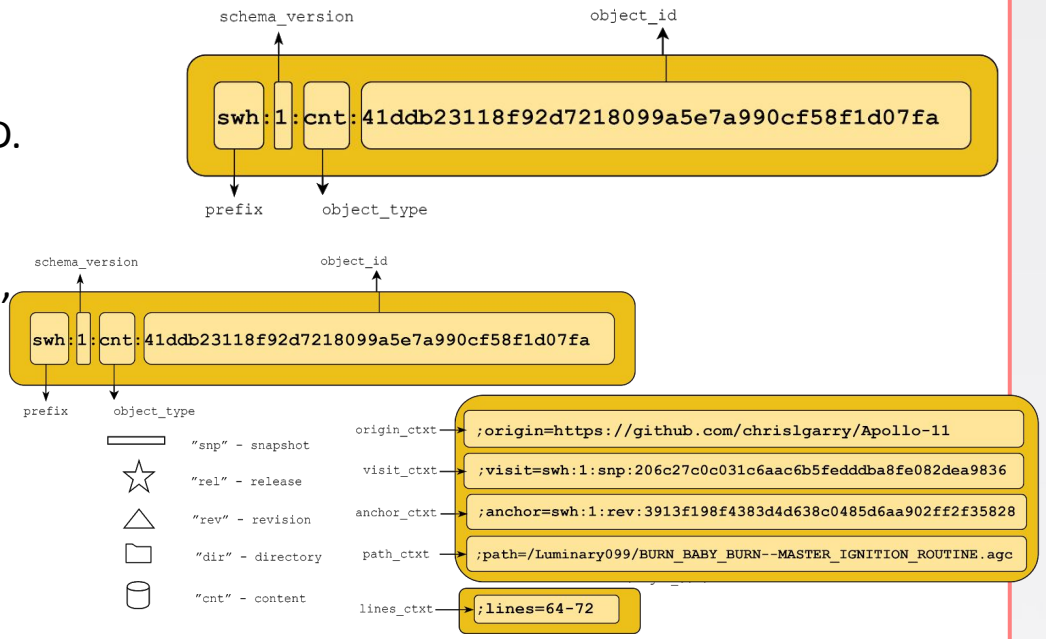
Intrinsic identifier

SWH provides a Persistent IDentifier (PID) that can identify each and every source code artifact with integrity, called a SWHID.

SWHIDs are intrinsic identifiers which are intimately bound to the designated object, they do not need a register, only agreement on a standard.

[Intrinsic vs. extrinsic blog post]

Go to API endpoint



Questions?
Demo

SWHID examples

Choose the granularity level for the reference:

code fragment

- swh:1:cnt:c60366bc03936eede6509b23307321faf1035e23;lines=473-537
- James McCaffrey's algorithm in sageMath

specific version - release

- swh:1:rel:22ece559cc7cc2364edc5e5593d63ae8bd229f9f
- release 2.3.0 of Darktable, dated 24 December 2016

full repository - snapshot

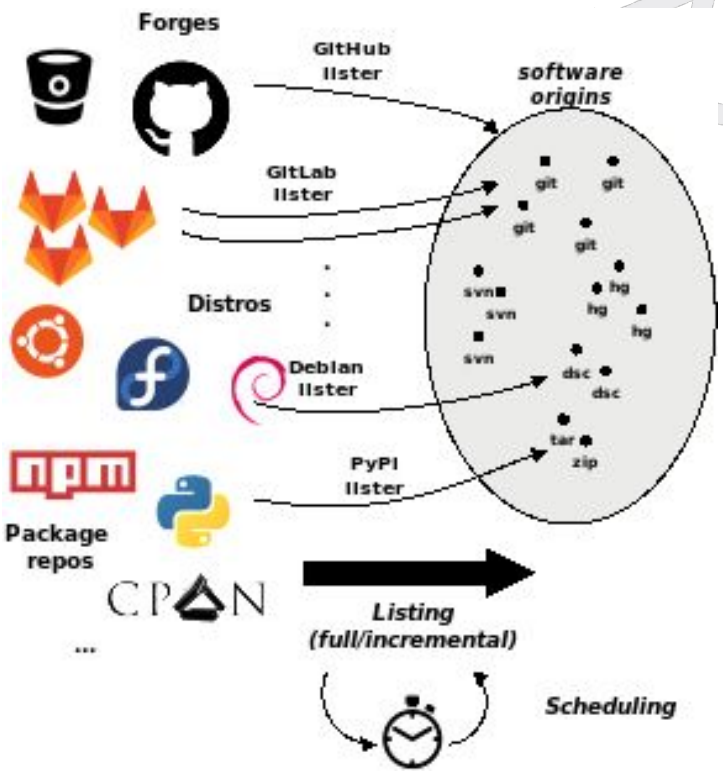
- swh:1:snp:c7c108084bc0bf3d81436bf980b46e98bd338453
- a snapshot of the entire Darktable repository (4 May 2017, GitHub)

Crawling

The SWH archive **harvests actively** source code from different sources and converts all the source code into a single and universal data structure which is an enormous Merkle DAG [Merkle, 1987].

Crawling is separated into three phases:

1. *listing software sources*,
2. *scheduling updates* and
3. *loading the software artifacts into the archive*.



Data model

The data model adopted by Software Heritage to represent the information that it collects is centered around the notion of *software artifact*, using the following canonical names, from bottom to top:

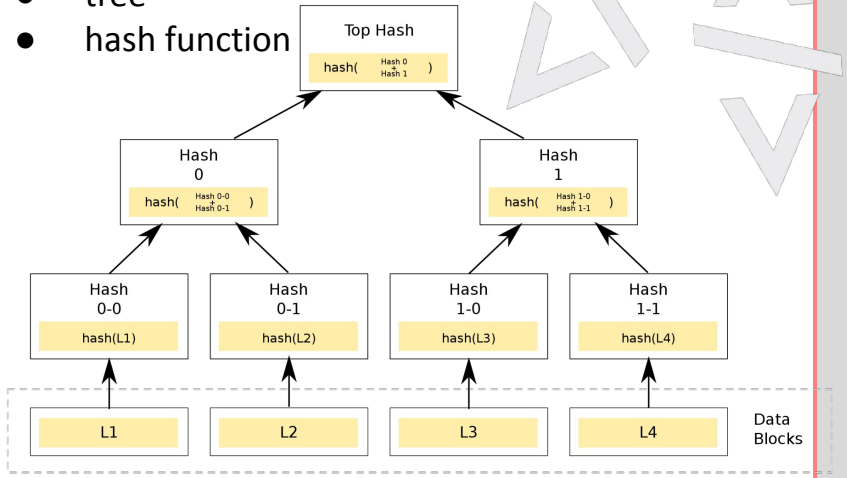
- contents,
- directories,
- revisions and
- releases.

Using also **origins**, **visits** and **snapshots** to store provenance information. Read more in [Software Heritage: Why and How to Preserve Software Source Code.]

Merkle tree

Combination of

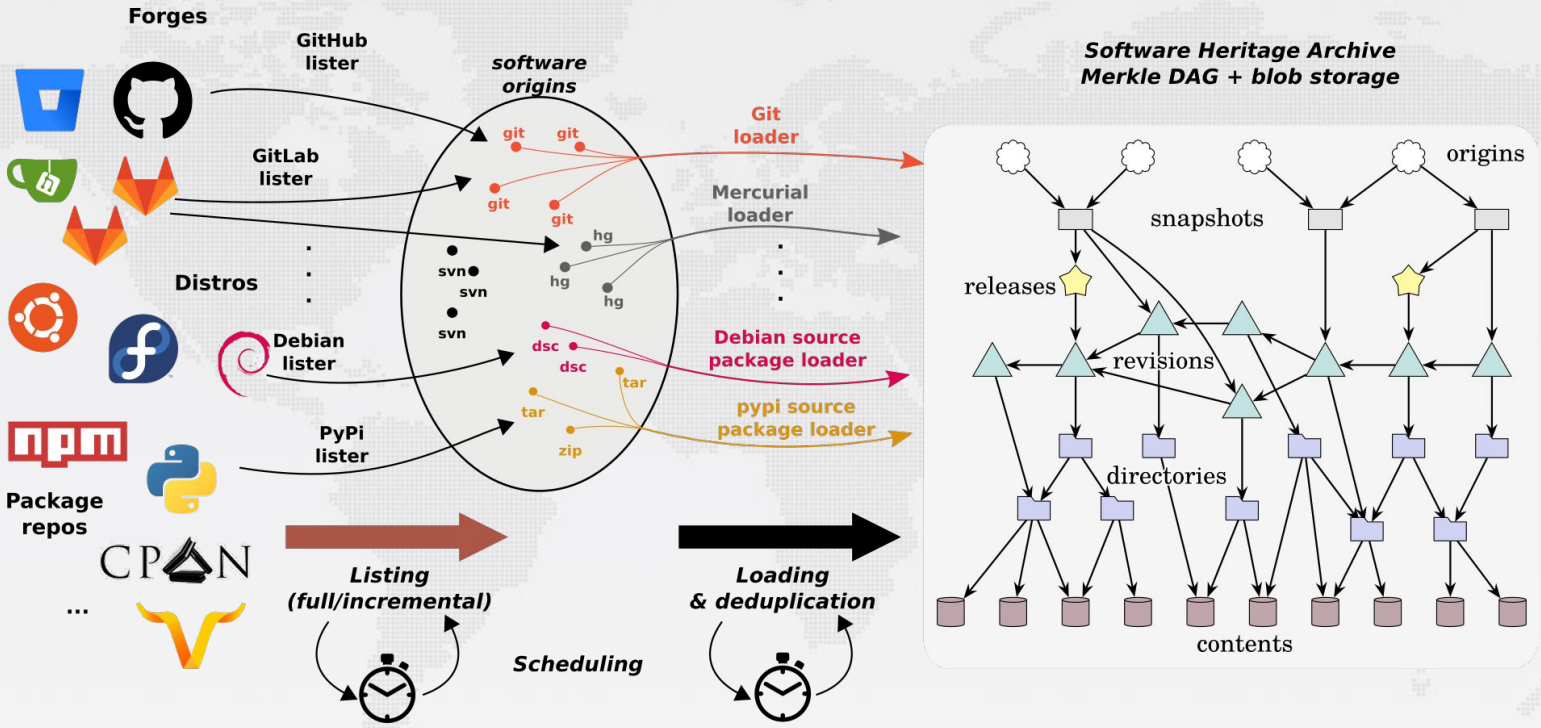
- tree
- hash function



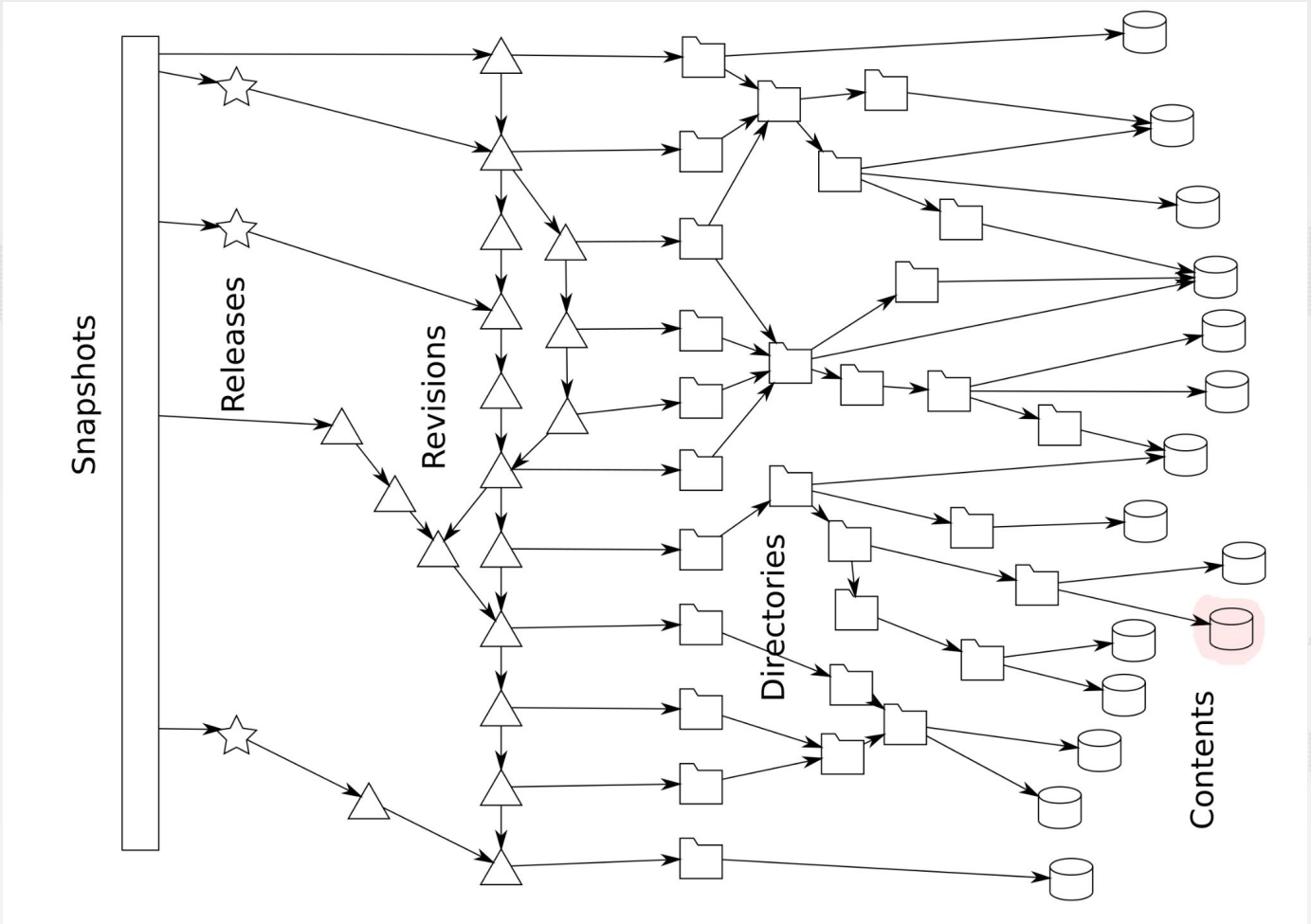
Classical **cryptographic construction**

- fast, parallel signature of large data structures
- widely used (e.g., Git, blockchains, IPFS, . . .)
- built-in deduplication

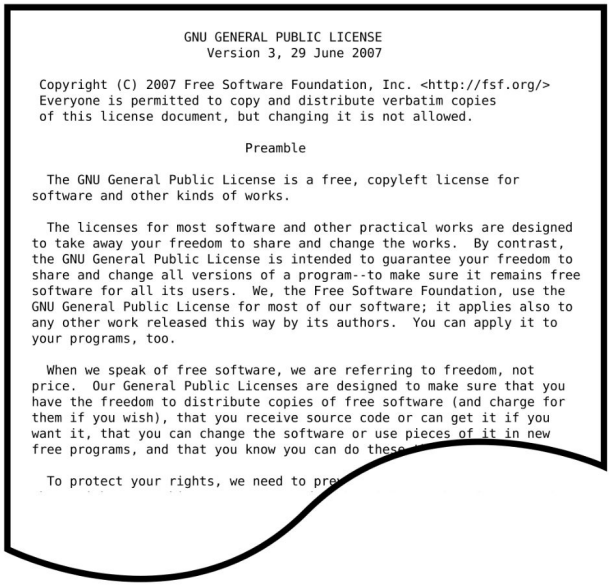
The complete workflow



Full development history permanently archived in a **uniform data model**


Snapshots
Releases
Revisions
Directories
Contents

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

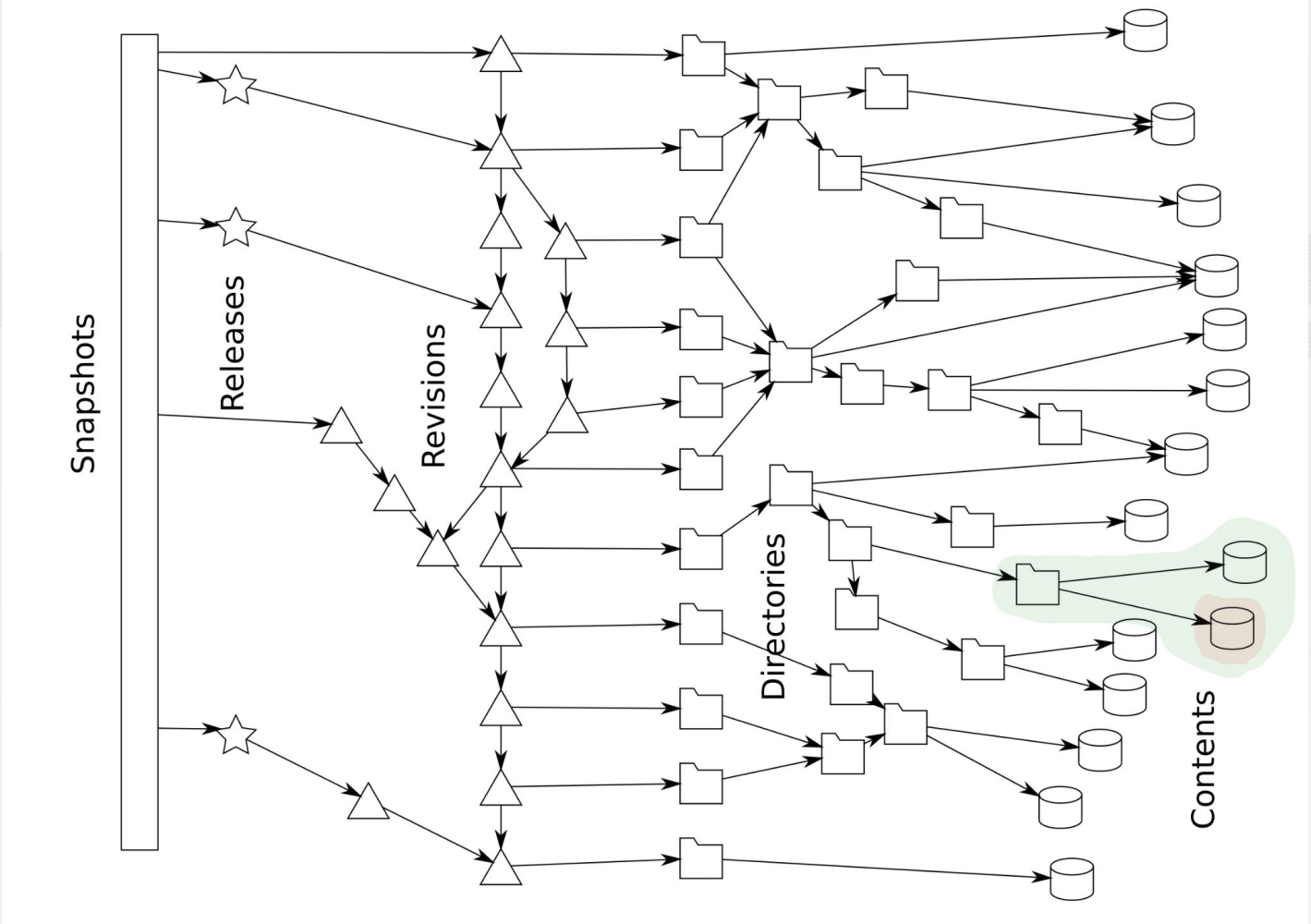
The GNU General Public License is a free, copyleft license for software and other kinds of works.

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

To protect your rights, we need to pre

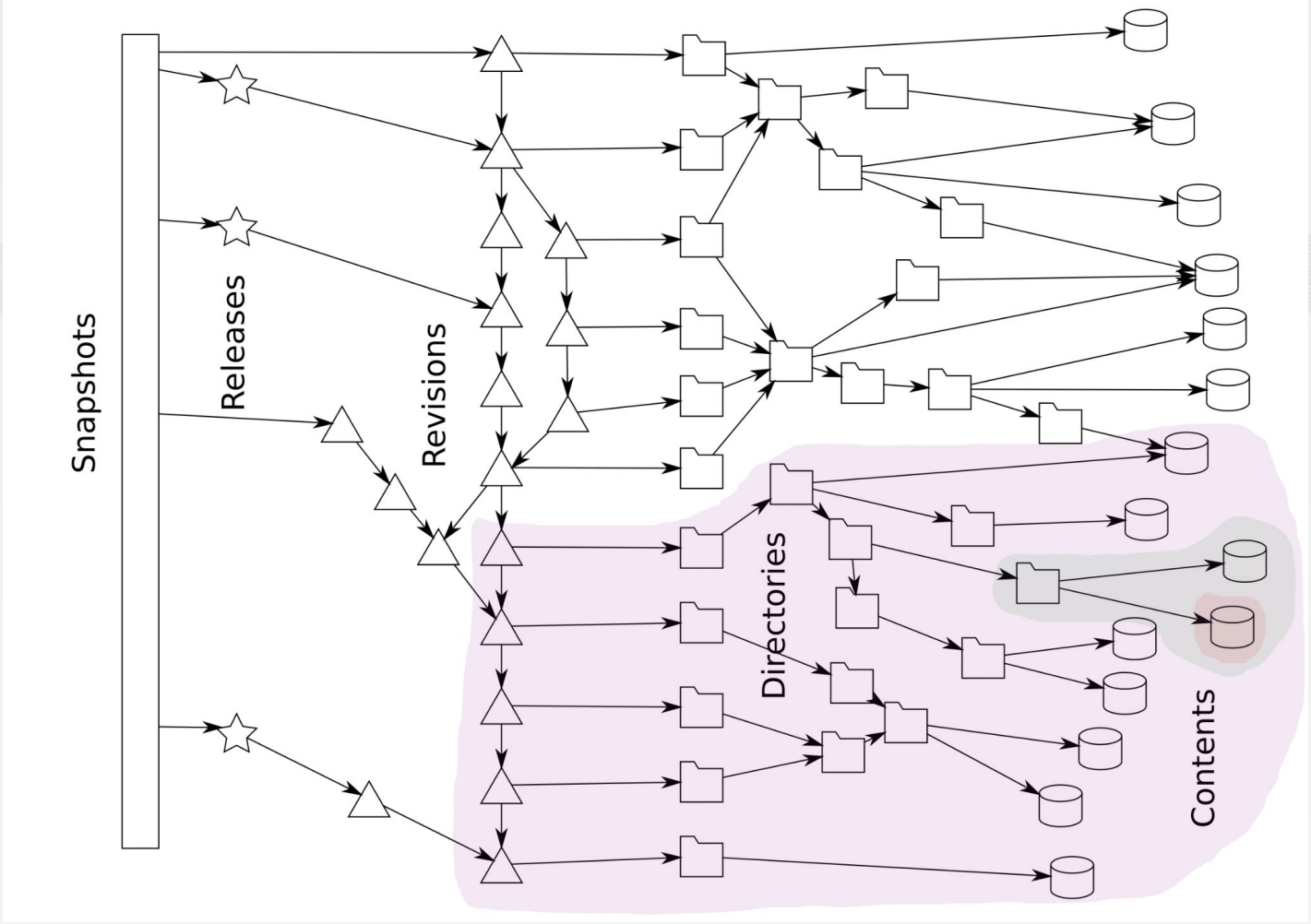
sha1: 8624bcdae55baeef...
sha256: 8ceb4b9ee5aded...
sha1_git: 94a9ed024d385...
length: 35147


Snapshots
Releases
Revisions
Directories
Contents

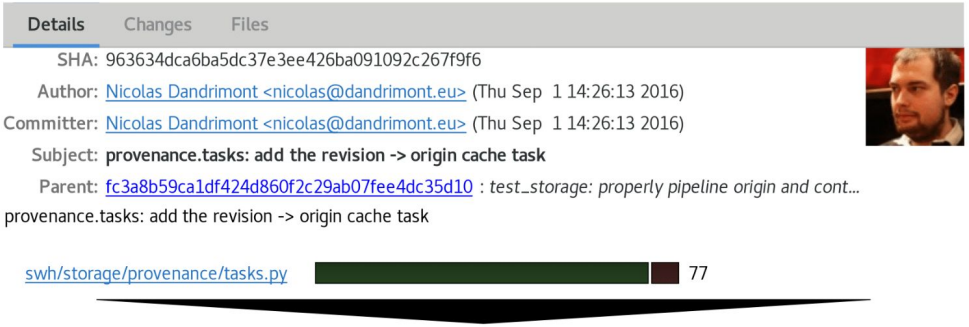
Directories

```
100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d0a34af6f52cf3cf6b0c2f7bd0648fbd255e77f AUTHORS
100644 blob 94a9ed024d3859793618152ea559a168bbcbb5e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e0867751dfb095c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5eaf2eb4 Makefile
100644 blob 5c7e3a5bbddb038682ba7793f440492ed9678bb3 Makefile.local
100644 blob 8617980629cd24e6080404f09aa749b085b3e07b README.db_testing
100644 blob 76b29f94cf815e0869c414d38d78d7ce08ec514e README.dev
040000 tree e1e10ecef948af0b93adb0372afc89f12e92618a bin
040000 tree 83e56d0beaf7793c77a45a345c80fcb8af503013 debian
040000 tree a34c9c4ba213f0cedc67f9816348d27955577af5 docs
100644 blob f2a6d32c6135aa7287bbd76167b01df2ae4f1539 requirements.txt
100755 blob eee147c36caf1bbc2d820da8dc026cb5b68180bc setup.py
040000 tree 224bb4c1f4c67fca1d160bffd2d06094e7e1abf3 sql
040000 tree 8631c9cd77bbe993168107ab5baf51f40c6300be swh
040000 tree 8fb905b56ba8ed692f1209b2773b474c6c1d66c1 utils
```

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d


Snapshots
Releases
Revisions
Directories
Contents

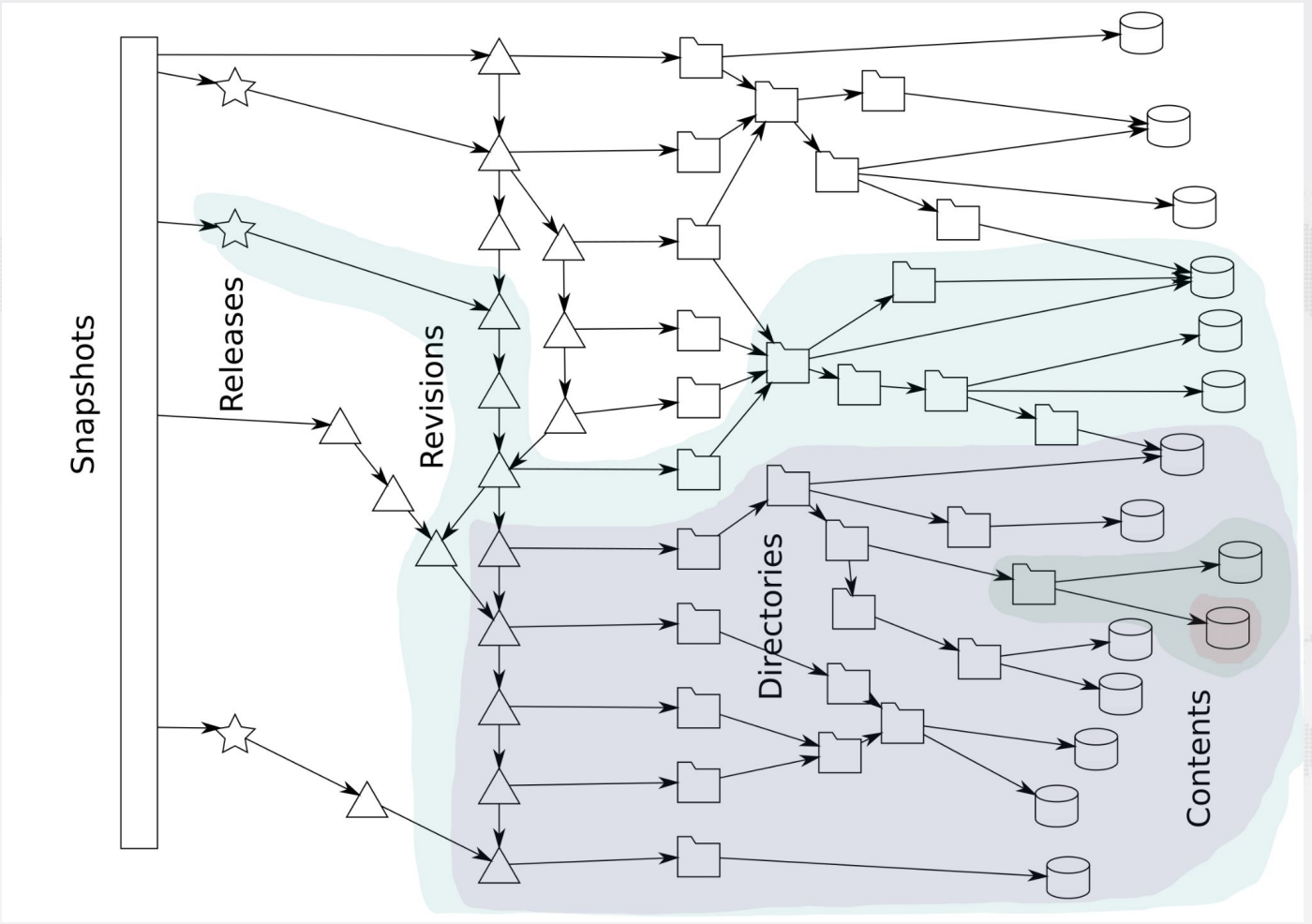
Revisions



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


Snapshots
Releases
Revisions
Directories
Contents

Releases

```
tag v0.0.51
Tagger: Nicolas Dandrimont <nicolas@dandrimont.eu>
Date:   Wed Aug 24 14:36:03 2016 +0200

Release swh.storage v0.0.51

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

commit c0c9f16b1e134f593e7567570a1761b156e6eb1d
```

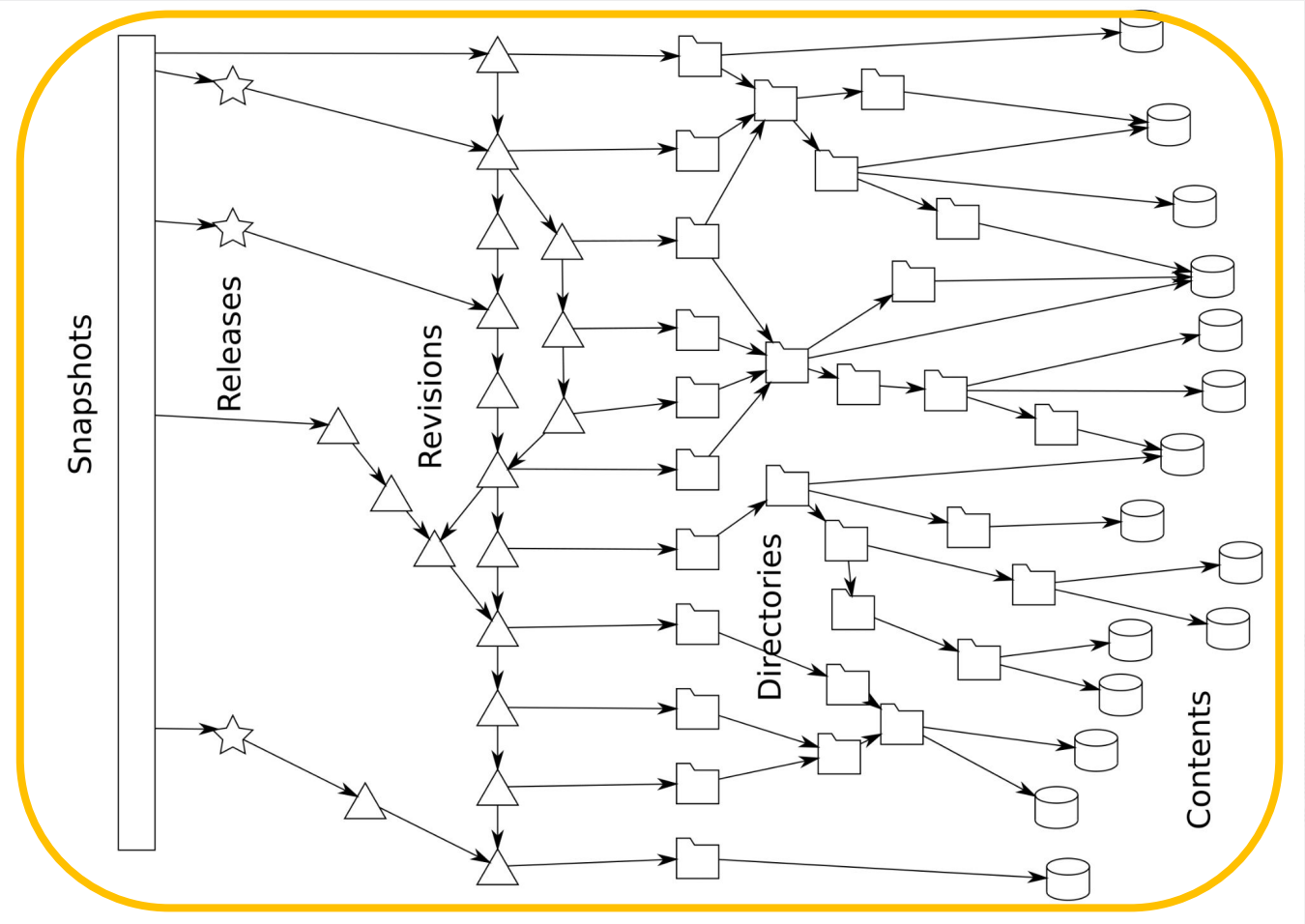
```
object c0c9f16b1e134f593e7567570a1761b156e6eb1d
type commit
tag v0.0.51
tagger Nicolas Dandrimont <nicolas@dandrimont.eu> 1472042163 +0200

Release swh.storage v0.0.51

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

iQIzBAABCAAdBQJXvZTNFhxuaWNvbGFzQGRhbmRyaW1vbnQuZXUACgkQ7AWLMo2+
neqorw//aq6SOb5DijzEa+kWN3rXgVS+1K1vEVh1wNKAwx8eKJ7aX2kEiLDtt7uf
ahpZ6pz3q8nqs6aC1+YrxBfcih3L2YtrdZeWXWqr8xWNMaEoYDb8qaphwh8AD5t2
lCBlit2uJtXuCrDt93eKKPwvzZXg+hB0sMWy35Dr6jW7Z7K4Mu/PGgIyIHPY55yo
lGEndWno7VfH1Vm6t1n5qB7l5mXRaqA+becqddubTZ2xjj+jplUqC8cyqN3hm/fL
qsj2mu8kyz3t8tG/H1/pV+l5OwBlnPoS5TH0tujoJEVgPK/dHSP79QuHDHZFkCao
klj6kAWyU80Mxb+nKV/jeLbrR3+yWBFj3Qp5a1/V8oOTh6E1dALcNMpEaKCoKtMt
d/gMRax1l1/g0EDfnsW67G6sDwKPKPHhgfVLQ3nV3GaQQTnu1RpMz006H9/tAwzC
Gg/K1PdHT4hzOiI46wYPZyje0U2VXGFu6vVU9vFQ4ZR/Wjn+0zMzdcRdrlJSUOMn
RpTTfUsbXUeXHGOpkgXhSYTnvp1gdPc76USTsK0aGe84AZm1lk0mGrwXCVfPqIYo
nhhibBSHBNMoqyF6yTSOpUbYK70tpYRRUGKWDeRK0wKSxkWKUZGtKzy6JYqIjo29
gulwgZQif5qWQCB0OontAL2+HvPFaVyckMejUhg62cP/+EHlvUk=
=kOxP
-----END PGP SIGNATURE-----
```

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb


Snapshots
Releases
Revisions
Directories
Contents

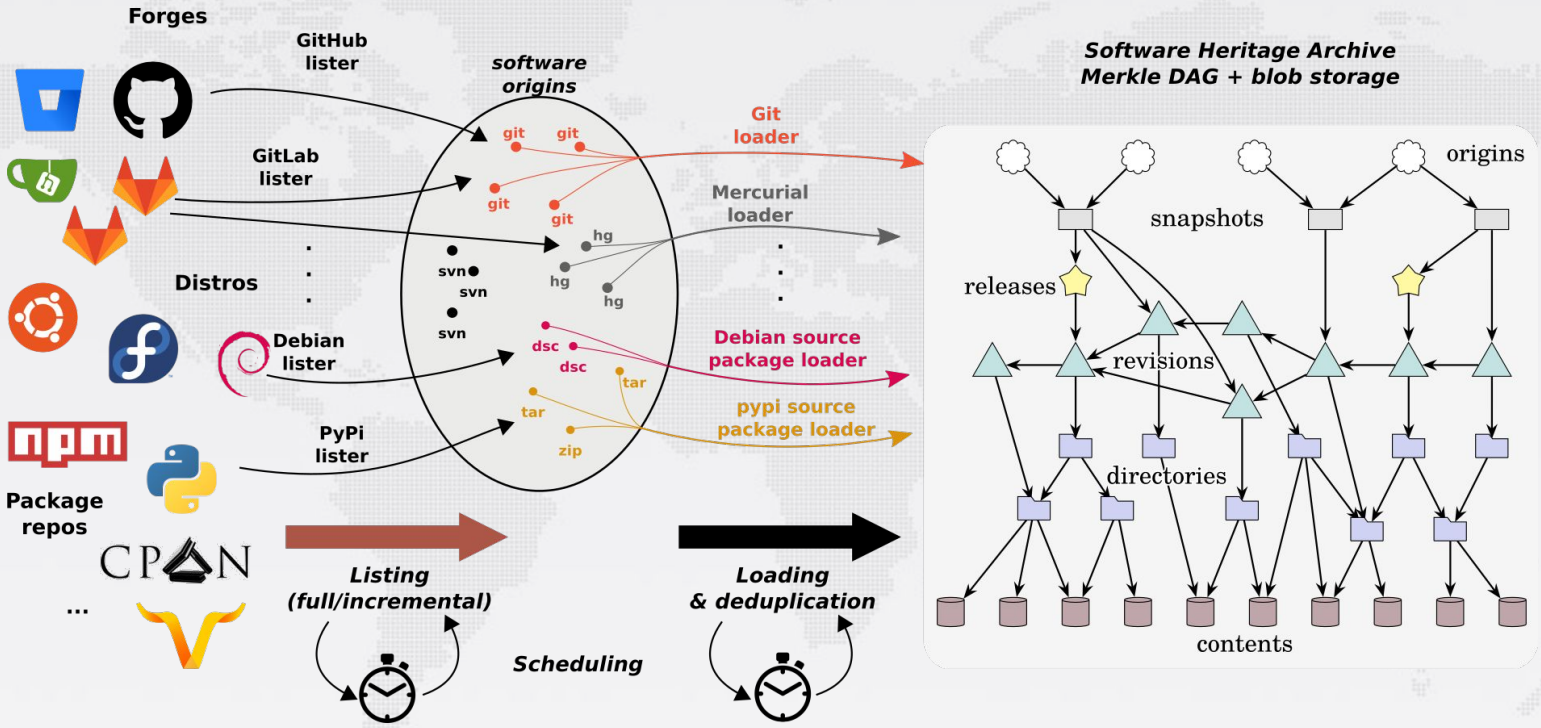
Snapshots

git show-refs

```
commit 08ffeb25770109525eb3ce21691466c53a1d9158 refs/heads/atime
commit ba5443a24e3f9fe323a46c292cec4fcbe61c67eb refs/heads/directory-listing-arrays
commit d69e0dbf892383ff6589b27fbe1c05d27238d9c5 refs/heads/foo
commit cf7ff9eea0eb22f8946908f5a8019f67de468e08 refs/heads/master
commit 7eca197fc66d2024047e54b1ed9e8b44361a0fc2 refs/heads/tmp-directory-add
commit 642a205f37de85005a85d427b53ee4fb2252e82e refs/heads/tmp/generic-releases
tag 20f043b1379cf768d966597799fd4907c757f755 refs/tags/v0.0.1
tag 72a21991a384e539996dbb867bfb0bee72aee2cd refs/tags/v0.0.10
tag 3590e0ca0ebb070e5b376705fa230bbfa4ffa5cc refs/tags/v0.0.11
tag 33378427a403ba569a67777b8d58f6674fbc6556 refs/tags/v0.0.12
tag 06f74652755b327cf590311c2bfa036cf3b4b35d refs/tags/v0.0.13
tag 5a6325fe86ab854b581d7442667d92a11e32f3bd refs/tags/v0.0.14
tag 586fba4e580b4f5fab05f599367643cbcb1a9c7f refs/tags/v0.0.15
tag 8cd8b885f4098bf363177742bd289f660e5be51c refs/tags/v0.0.16
tag a542444ee3f0fbed35efb202fee035c809abc7d6 refs/tags/v0.0.17
tag 228a2f1650dd12222e556559462e1e06fc4993d9 refs/tags/v0.0.18
tag 606979a4ca05d497fc0d24aad00dce82636ef47c refs/tags/v0.0.19
tag 32bf5a59fc2a323baa6d5f15a6ad5382ec275a67 refs/tags/v0.0.2
tag 3147c3d31ec46cf6492f881e908b1237ebdff2c7 refs/tags/v0.0.20
tag 215ea50daba111e082e0b72e76eb4b6073a87908 refs/tags/v0.0.21
tag 3fb168c2072a5d6252124257a1e5dfc0f5ffa1df refs/tags/v0.0.22
tag 8cdbee8da4d73fc5d262789e460a16ac3c72aba4 refs/tags/v0.0.23
...
```

id: b464cad1b66fff266a37b46ea6e7a04b545e904b

The complete workflow



Full development history permanently archived in a **uniform data model**

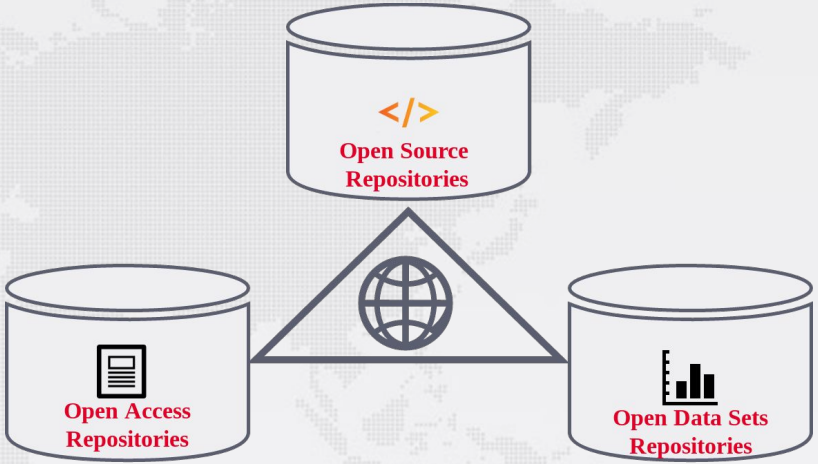
Outline

- ★ Introduction
- ★ Preserving source code
- ★ Software Heritage the universal source code archive
- ★ **Research Software: a first class research output**
- ★ Good practices for software curation
- ★ Conclusion

Software in Research: A pillar of Open Science

Multiple facets, it can be seen as:

- a **tool**
- a research **outcome** or result
- **the object** of research



Three pillars of Open Science
Gruenpeter, Software Heritage CC-By 4.0 2019

Much more complex than it seems

- **Structure**
  - monolithic/
  - composite;
  - self-contained/
  - external dependencies
- **Lifetime**
  - One-shot / long term
- **Community**
  - one person / one team/distributed community
- **Authorship**
  - complex set of roles
- **Authority**
  - institutions/organizations/communities/single person

Various granularities

- **Exact status** of the source code for **reproducibility**, e.g.
  - “you can find at swh:1:cnt:cdf19c4487c43c76f3612557d4dc61f9131790a4;lines=146-187 the core algorithm used in this article”
- (Major) **release** “This functionality is available in OCaml **version 4**”
- **Project** “Inria has created OCaml and Scikit-Learn”

A plurality of needs

Researchers

- **archive and reference** software used and created in articles
- **find** useful software
- **get credit** for developed software
- **verify/reproduce/improve** results

Research Organization

know its **software assets** for:

- technology transfer,
- impact metrics,
- strategy

Laboratories/teams

- **track** software contributions
- **produce** reports
- **maintain** web page

Open science/source in France

L'ouverture des codes sources des logiciels est un enjeu majeur de reproductibilité des résultats scientifiques

Définir et promouvoir une politique en matière de logiciels libres

- Établir une **Charte nationale des logiciels libres** issus de l'enseignement supérieur, de la recherche et de l'innovation.
- Développer le lien entre données et logiciels grâce au réseau des **administrateurs des données, des algorithmes et des codes sources** dans les établissements.
- Émettre des **recommandations auprès des organismes financeurs** pour accompagner au mieux le développement logiciel.
- Faire monter en compétence les structures de valorisation sur **les modèles économiques associés** à la production de logiciels libres.
- Soutenir **Software Heritage** et recommander son adoption pour l'archivage et le référencement des codes sources.

Reconnaître les codes sources comme une contribution à la recherche

- Créer un **prix du logiciel libre pour la recherche** qui récompense les équipes et projets exemplaires dans le domaine.
- **Mieux valoriser** les productions logicielles dans la carrière des chercheurs, des personnels d'accompagnement à la recherche et dans l'évaluation des structures de recherche.
- **Suivre dans le temps** la production de codes et logiciels de la recherche française pour en identifier les dynamiques, l'ouverture et les impacts grâce au baromètre de la science ouverte.
- Construire un **catalogue** des logiciels issus de la recherche en utilisant un schéma de métadonnées normalisé et partagé entre tous les acteurs de l'enseignement supérieur, de la recherche et de l'innovation.

Source: Ministère de l'enseignement supérieur et de la recherche

Outline

- ★ Introduction
- ★ Preserving source code
- ★ Software Heritage the universal source code archive
- ★ Research Software: a first class research output
- ★ **Good practices for software curation**
- ★ Conclusion

What is at stake? In order of difficulty

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 them (visibility)

Cite (for credit)

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

➔ to give credit to authors (evaluation!)

Archive

Submit code to HAL

- A scholarly repository
- An archive
- Software is more finadable
- Transferring code to SWH

Save code now to SWH

- Easy (only submit repo's url)
- All dev history archived
- different vcs are accepted
- PID to reference specific pieces of code (even algorithms)

Classic deposit

.zip, .tar.gz

SWHID deposit

Save code now

Git, svn, hg

HAL and SWH collaboration

Key dates

- ★ 2017 - Collaboration launch
- ★ Mars 2018 - Beta-test on HAL-Inria
- ★ Septembre 2018 - Launch on all HAL instances
- ★ April 2020 - BibLaTeX @software export
- ★ April 2022 - Deposit with SWHID (beta test on HAL-Inria)
  - Release on YouTube of the Open Science tutorial series: software source code deposit

The actors

HAL open science

CCSD

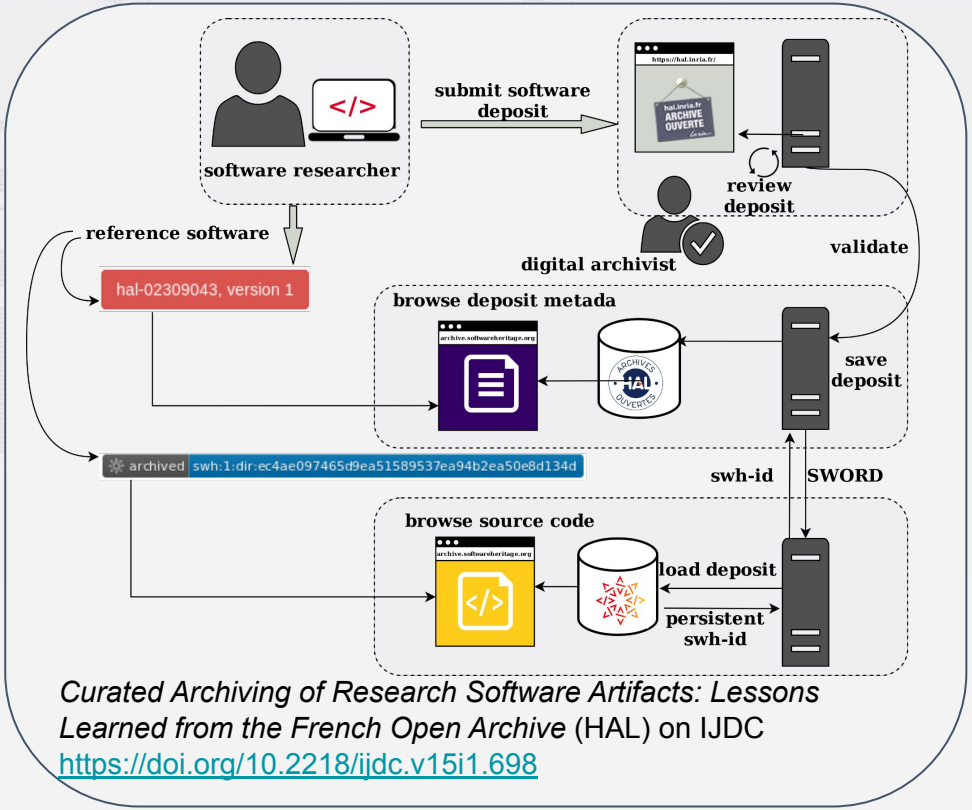
Inria

Software Heritage

Classic deposit

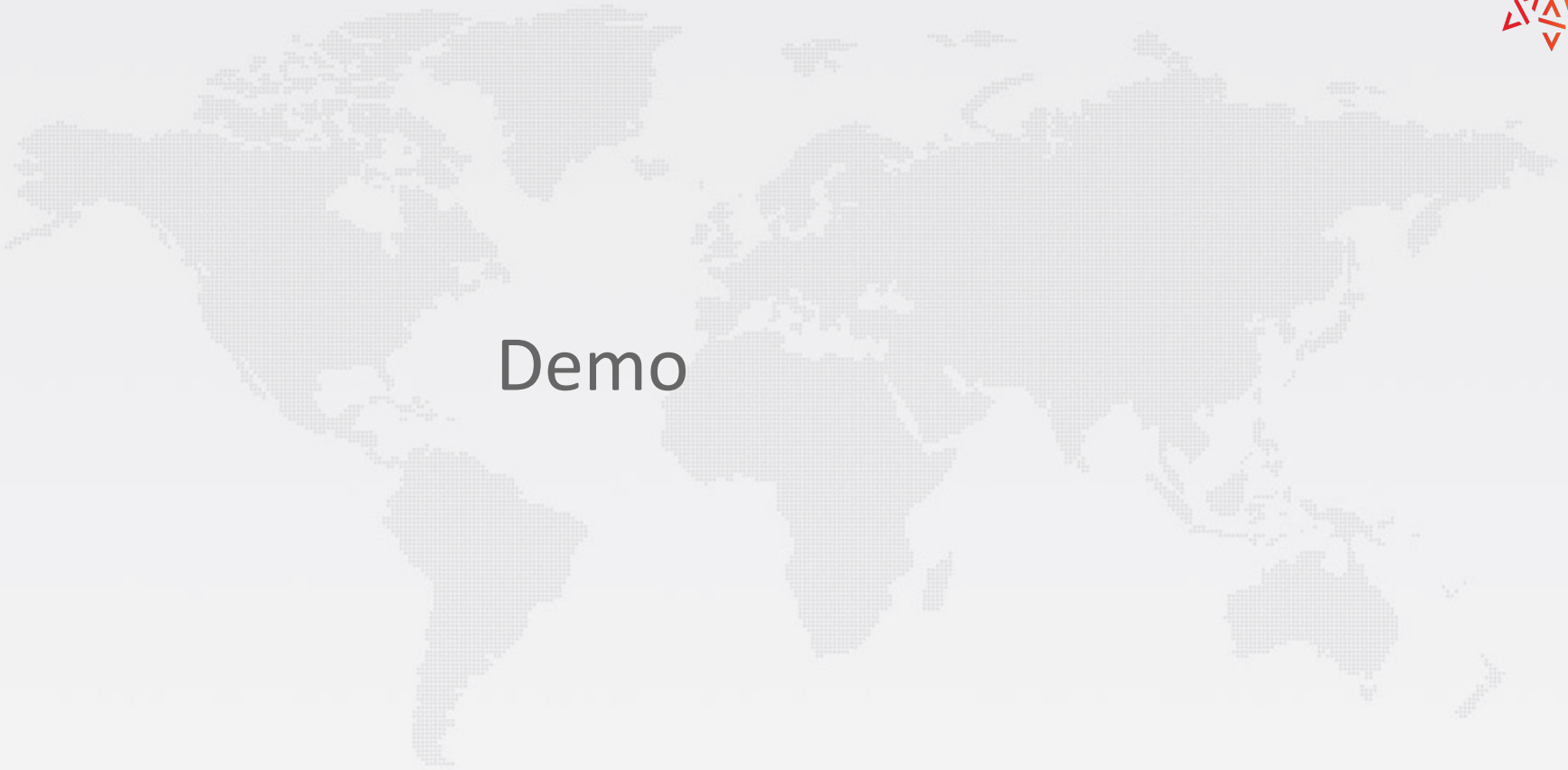
Advantages:

- ★ Metadata moderation by the digital archivist team
- ★ Export formats available on the software record - to cite software



Curated Archiving of Research Software Artifacts: Lessons Learned from the French Open Archive (HAL) on IJDC
https://doi.org/10.2218/ijdc.v15i1.698

Deposit guide :
Morane Gruenpeter, Jozefina Sadowska, Estelle Nivault, Alain Monteil. Create software deposit in HAL: User guide and best practices. [Technical Report] Inria; CCSD; Software Heritage. 2022. ⟨hal-01872189v2⟩

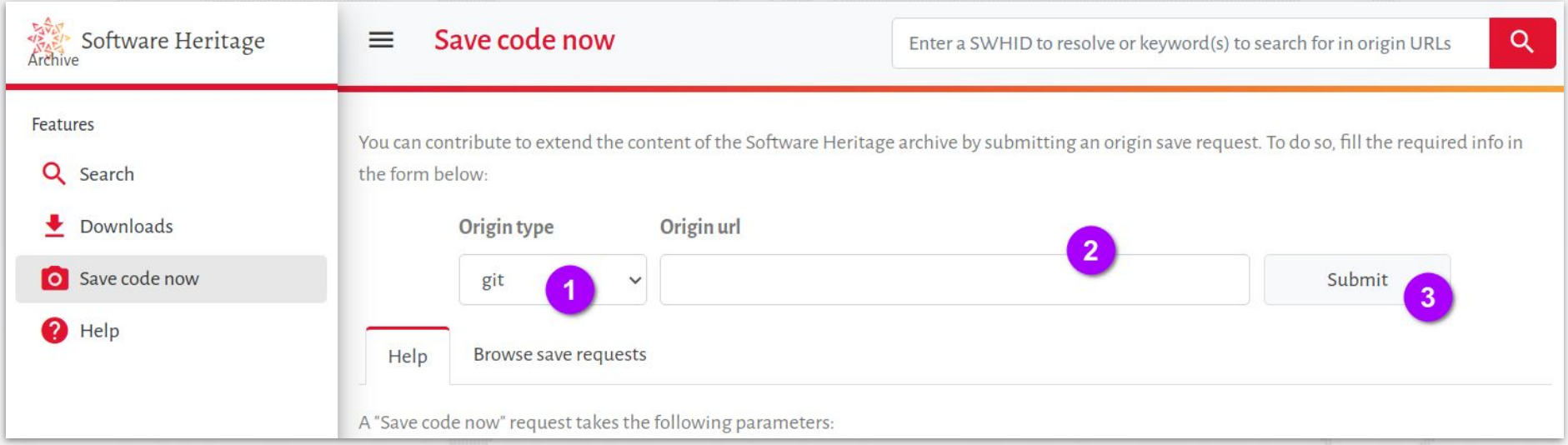


Demo

Save ~~your~~ any code now!

https://save.softwareheritage.org/

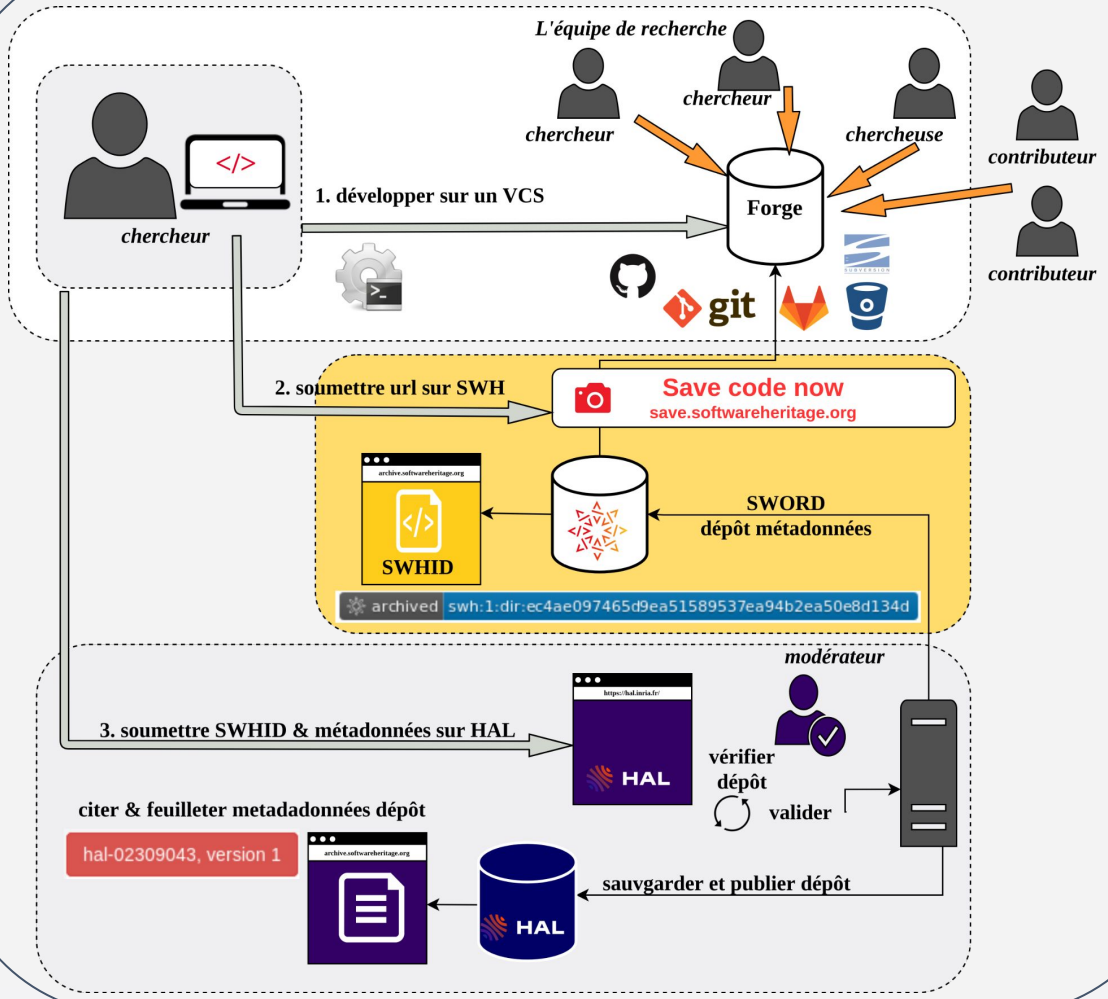
Save code now

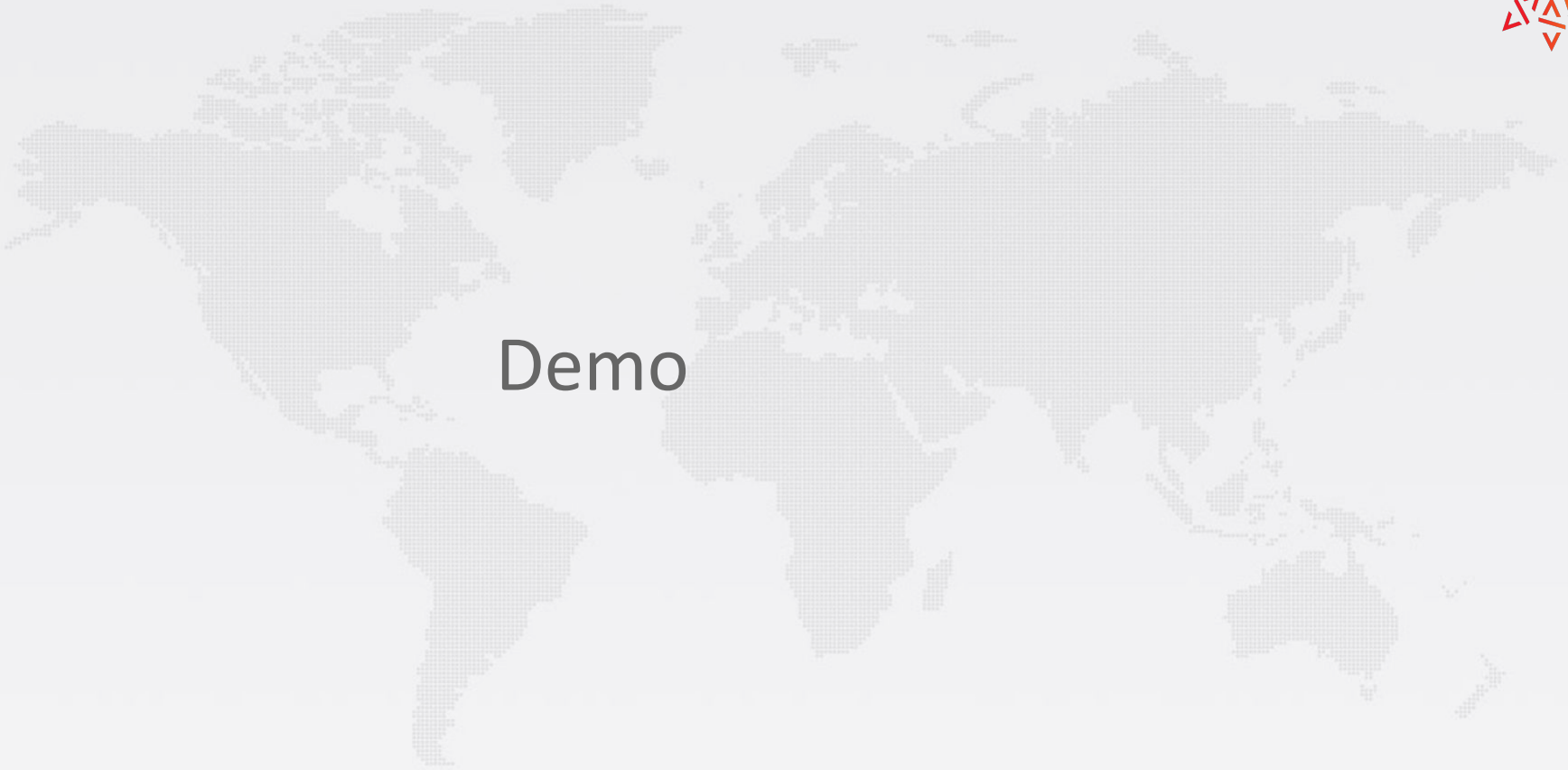


SWHID deposit

Advantages:

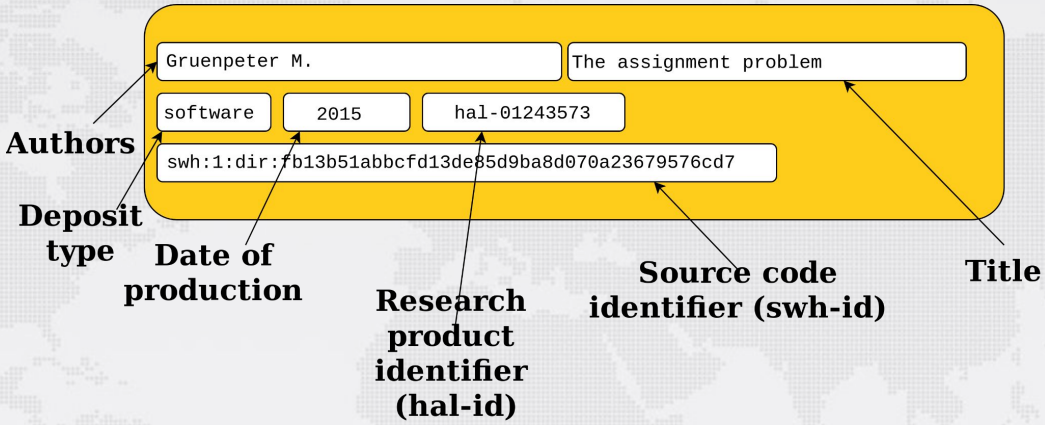
- ★ Metadata moderation by the digital archivist team
- ★ Export formats available on the software record - to cite software
- ★ Retrieve metadata with codemeta.json to complete form





Demo

Reference vs. citation



Credit & Attribution

a metadata record
all authors & contributors

Reuse & Reproducibility

a specific artifact
with complementary information (docs)

Archive & Index

metadata record (HAL)
artifact itself (SWH)

Citation export in HAL

- ➔ Citation accessible on the HAL record
- ➔ Export BibTeX using the format BibLaTeX **@software** or **@softwareversion** (if a version property was submitted)
- ➔ Export used in activity reports for scientific outputs at Inria since 2020.

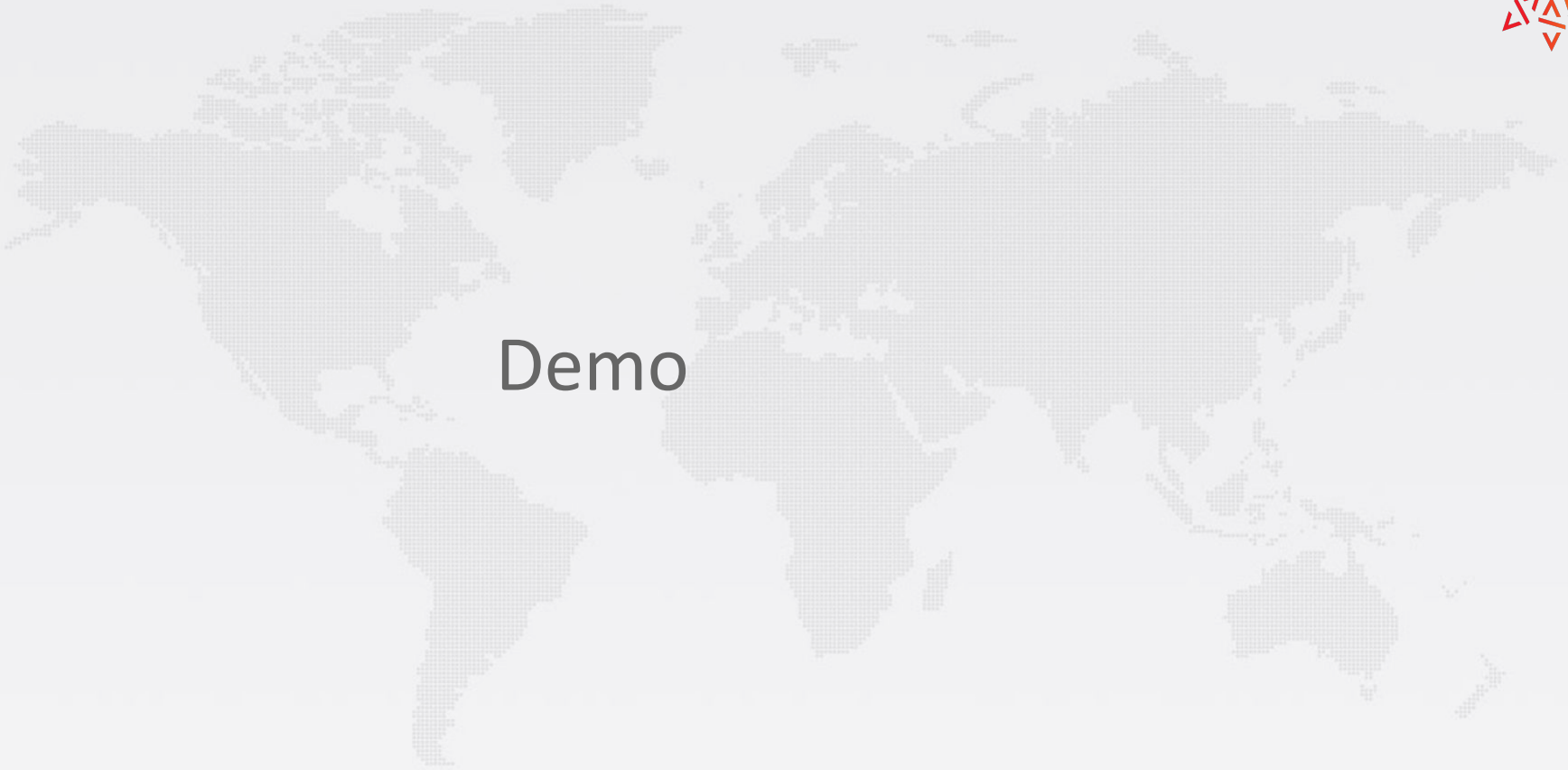
HAL's citation format

Matteo Frigo, Mauro Zucchelli, Rachid Deriche, Samuel Deslauriers-Gauthier. TALON: Tractograms As Linear Operators in Neuroimaging. 2021, ⟨swh:1:dir:f25157ad1b13cb20ac3457d4f6756b49ac63d079;origin=https://hal.archives-ouvertes.fr/hal-03116143;visit=swh:1:snp:465d8995619657871 7f4cb5155e456c279aa6a22;anchor=swh:1:rev:10247a14640a280b9140a27ce003d382d70cccac;path=/⟩. ⟨hal-03116143⟩

```
@softwareversion{frigo:hal-03116143v1,
  TITLE = {{TALON: Tractograms As Linear Operators in
Neuroimaging}},
  AUTHOR = {Frigo, Matteo and Zucchelli, Mauro and
Deriche, Rachid and Deslauriers-Gauthier, Samuel},
  URL = {https://hal.archives-ouvertes.fr/hal-03116143},
  NOTE = {},
  YEAR = {2021},
  MONTH = Jan,
  SWHID =
{swh:1:dir:f25157ad1b13cb20ac3457d4f6756b49ac63d079;origin
=https://hal.archives-ouvertes.fr/hal-03116143;visit=swh:1
:snp:465d89956196578717f4cb5155e456c279aa6a22;anchor=swh:1
:rev:10247a14640a280b9140a27ce003d382d70cccac;path=/},
  VERSION = {0.3.0},
  REPOSITORY = {https://gitlab.inria.fr/cobcom/talon},
  LICENSE = {MIT License},
  KEYWORDS = {diffusion MRI ; dMRI ; tractography ; python
; optimization},
  FILE =
{https://hal.archives-ouvertes.fr/hal-03116143/file/talon-
source.zip},
  HAL_ID = {hal-03116143},
  HAL_VERSION = {v1},
}
```

Softwares

[38] [SW] M. Frigo, M. Zucchelli, R. Deriche and S. Deslauriers-Gauthier, *TALON: Tractograms As Linear Operators in Neuroimaging* version 0.3.0, 19th Jan. 2021. HAL: ⟨hal-03116143⟩, URL: https://hal.archives-ouvertes.fr/hal-03116143, VCS: https://gitlab.inria.fr/cobcom/talon, SWHID: ⟨swh:1:dir:f25157ad1b13cb20ac3457d4f6756b49ac63d079;origin=https://hal.archives-ouvertes.fr/hal-03116143;visit=swh:1:snp:465d89956196578717f4cb5155e456c279aa6a22;anchor=swh:1:rev:10247a14640a280b9140a27ce003d382d70cccac;path=/⟩.



Demo

Citation with the biblatex-software package

BibLaTex style extension for software

Youtube tutorial

[Software Release] B. Langmead and S. L. Salzberg, *Bowtie2* version 2.4.2, Oct. 2022. LIC: GPL. URL: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml, VCS: https://github.com/BenLangmead/bowtie2, SWHID: ⟨swh:1:rel:97bacffeaa6e7c3f574ce5b566daba82aa18a11f;origin=https://github.com/BenLangmead/bowtie2;visit=swh:1:snp:c25778cfefc086c63c6f78eed230d0b9c88876ee⟩.

[Software excerpt] MIT Instrumentation Laboratory, "AGC Luminary routine for changing LEM asset during landing", from *Apollo 11 Guidance Computer (AGC) source code for the command and lunar module* 1967. VirtualAGC project. LIC: Public Domain. URL: https://www.ibiblio.org/apollo, VCS: https://github.com/virtualagc/virtualagc, SWHID: ⟨swh:1:cnt:64582b78792cd6c2d67d35da5a11bb80886a6409;origin=https://github.com/virtualagc/virtualagc;anchor=swh:1:rev:007c2b95f301f9438b8b74d7993b7a3b9a66255b;lines=245-261⟩.

Describe - mandatory files for a HAL deposit

★ Prepare your code with the following files :

*These files are verified by **moderators***

- ❑ README (https://readme.so/fr/editor)
- ❑ AUTHORS (containing list of authors)
- ❑ LICENSE
  - Open-source SPDX compliant license
    https://choosealicense.com/
    https://reuse.software/
- ❑ codemeta.json (not mandatory but useful)

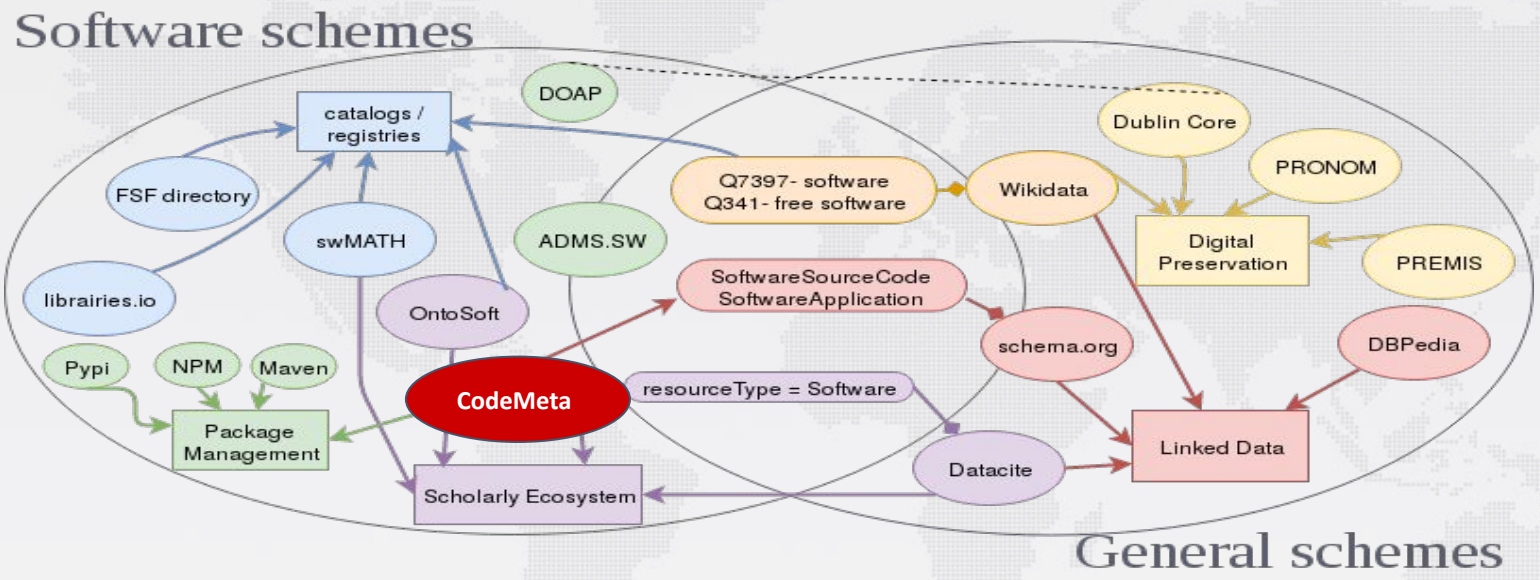
See also: HOWTO archive and reference your code

Describe: What's a good README

- ★ **MUST** include:
  - Name and a description of the software.
- ★ **SHOULD** include:
  - how to run and use the source code
  - build environment, installation, requirements
- ★ **CAN** include:
  - project website or documentation pointer and recent news
  - visuals

extracted from Eric Steven Raymond and Make a README

Describe - choose a vocabulary?



Software ontologies landscape from Pathways for Discovery of Free Software (slide deck from LibrePlanet 2018). (Gruenpeter & Thornton, 2018) CC-by-4

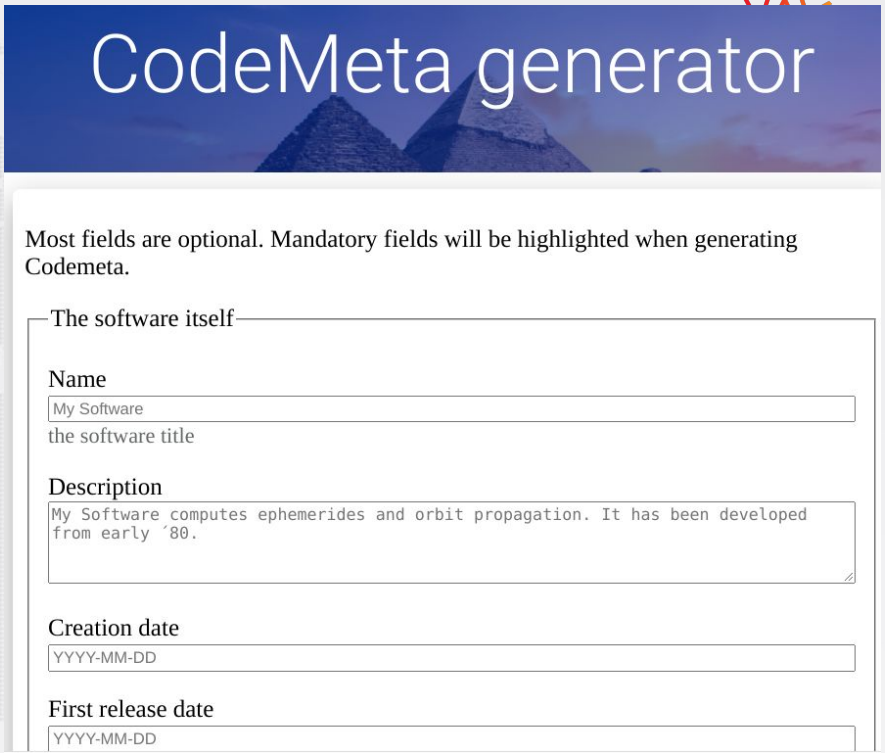
CodeMeta

- An initiative
  - An academic community discussing software metadata
- A vocabulary
  - A subset of schema.org
- A crosswalk table - mapping the metadata landscape

An open source tool to create codemeta.json files

Contributed to the community by

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE



Use it directly on the CodeMeta hosted version

Contributions are welcome on the code repository

Describe: Software metadata terms

Identify

- identifier
- name
- author(s)
- contributor(s)
- version
- applicationCategory
- codeRepository

administrate

- maintainer
- citation
- funder(s)
- license
- editor
- publisher
- dates
  - created,
  - modified
  - published
- developmentStatus

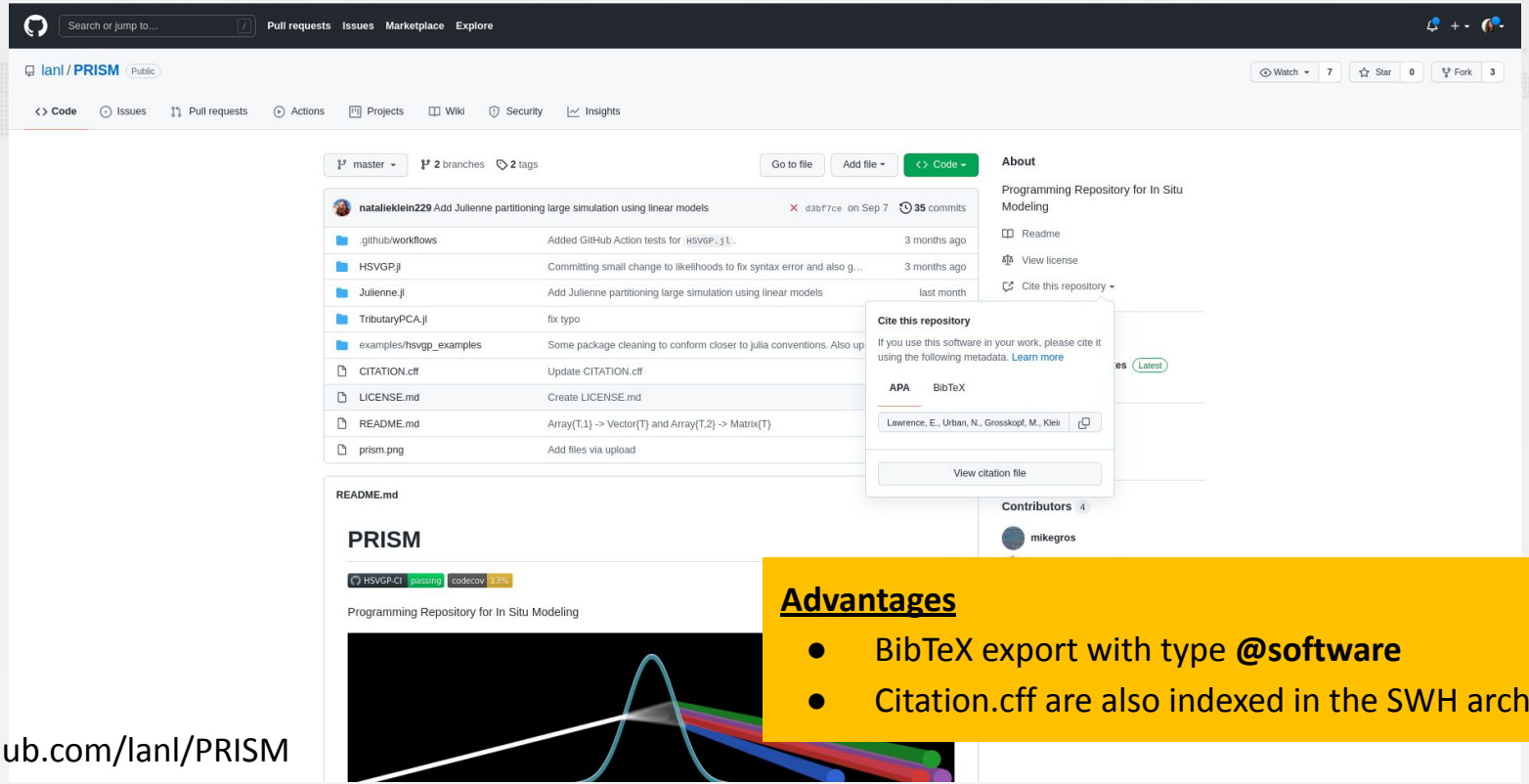
classify

- description
- keywords
- supportingData
- referencePublication
- algorithms*
- readme (docs*)

execute

- buildInstructions
- issueTracker
- operatingSystem
- softwareRequirements
- runtimePlatform
- downloadUrl
- memory, procesor, storage

Another perspective: CITATION.cff



Advantages

- BibTeX export with type **@software**
- Citation.cff are also indexed in the SWH archive

Source:
https://github.com/lanl/PRISM

Codemeta.json vs. CITATION.cff

Branch: **master** ▾ **doi2cff** / **CITATION.cff**

sverhoeven Added DOI badge and cff

1 contributor

21 lines (20 sloc) | 791 Bytes

```
# YAML 1.2
# Metadata for citation of this software according to the CFF fo
cff-version: 1.0.3
message: If you use this software, please cite it as below.
# FIXME title as repository name might not be the best name, ple
title: DOI 2 citation format file generator
doi: 10.5281/zenodo.1206049
# FIXME splitting of full names is error prone, please check if
authors:
- given-names: Stefan
  family-names: Verhoeven
  affiliation: Netherlands eScience Center
- given-names: Jurriaan
  family-names: Spaaks
  name-particle: H.
  affiliation: Netherlands eScience Center
version: 1.0.0
date-released: 2018-03-23
repository-code: https://github.com/citation-file-format/doi2cff
license: Apache-2.0
```

Fichier de citation utilisé pour le logiciel doi2cff: https://github.com/citation-file-format/doi2cff/blob/master/CITATION.cff

Branch: **master** ▾ **CMM** / **codemeta.json**

scatenag Update codemeta.json

1 contributor

93 lines (90 sloc) | 4.28 KB

```
{
    "@context": "https://doi.org/10.5063/schema/codemeta-2.0",
    "@type": "SoftwareSourceCode",
    "identifier": "CMM",
    "description": "conservative garbage collector for C++ develope
    "name": "Customizable Memory Management",
    "codeRepository": "https://github.com/Unipisa/CMM",
    "applicationCategory" : "Memory Management",
    "license": "  Copyright (c) 1993, 1994, 1995 Giuseppe Attardi a
    "version": "1.9",
    "keywords" : [
      "Garbage Collector",
      "Memory Management"
    ],
    "runtimePlatform" : "SunOS 4.x, Solaris 2.x, Linux 1.x, 2.x, AI
    "softwareRequirements" : ["g++"],
    "developmentStatus" : "Unsupported",
    "dateCreated" : "1994-08-27",
    "datePublished" : "1995-08-26",
        "dateModified": "1998-03-03",

```

Outline

- ★ Introduction
- ★ Preserving source code
- ★ Software Heritage the universal source code archive
- ★ Research Software: a first class research output
- ★ Good practices for software curation
- ★ **Conclusion**

Wrap up

- 📦 Archive your code!

  https://save.softwareheritage.org/

- 👓 Describe your code with metadata

  README, LICENSE, AUTHORS, codemeta.json

- 🗺️ Reference your code

  SWHID over DOI, context

- 🧑‍🎓 Cite your code

  Version, release, file, lines

The SWH ambassadors program

Who can be an ambassador?

- ★ **Enthusiastic individuals** who wish to volunteer as ambassadors to help grow the Software Heritage community
- ★ An ambassador can come from **different areas**; academia, cultural heritage, industry and public administration.

Do I need a technical background to become an ambassador?

- ★ Absolutely not! Source code archival is a cross-domain, cross-expertise concern.
- ★ you can have a technical background which might be helpful to read technical documentation and trying some of the more advanced features.

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Inria

INVENTEURS DU MONDE NUMÉRIQUE

Thank you for your attention!

Keep in touch: morane@softwareheritage.org
@moraneottilia, @SWHeritage
https://www.softwareheritage.org/newsletter/

alexis.lebis@imt-nord-europe.fr