Software Heritage

A research platform for large-scale source code archival

Stefano Zacchiroli

Télécom Paris — zack@upsilon.cc, @zacchiro

19 Oct 2021
IDIA Prototype/Software/Platform Day
Télécom Paris

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

About me

- Professor, Télécom Paris, ACES team (newbie!)
- Free/Open Source Software activist (20+ years)
- Debian Developer & Former 3x Debian Project Leader
- Former Open Source Initiative (OSI) director
- Software Heritage co-founder & CTO

Research interests

- Software engineering, Free/Open Source Software (FOSS)
- Digital commons
- Computer security
- Software supply chain

Outline

Software *source code* is precious knowledge

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

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

Software *source code* is precious knowledge

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

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

Apollo 11 source code (excerpt)

```
P63SPOT3        CA      BIT6            # IS THE LR ANTENNA IN POSITION 1 YET
                EXTEND
                RAND    CHAN33
                EXTEND
                BZF     P63SPOT4        # BRANCH IF ANTENNA ALREADY IN POSITION 1

                CAF     CODE500         # ASTRONAUT:    PLEASE CRANK THE
                TC      BANKCALL        #               SILLY THING AROUND
                CADR    GOPERF1
                TCF     GOTOPOOH        # TERMINATE
                TCF     P63SPOT3        # PROCEED       SEE IF HE'S LYING

P63SPOT4        TC      BANKCALL        # ENTER         INITIALIZE LANDING RADAR
                CADR    SETPOS1

                TC      POSTJUMP        # OFF TO SEE THE WIZARD ...
                CADR    BURNBABY
```

Software *source code* is precious knowledge

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

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

Apollo 11 source code (excerpt)

```
P63SPOT3        CA      BIT6            # IS THE LR ANTENNA IN POSITION 1 YET
                EXTEND
                RAND    CHAN33
                EXTEND
                BZF     P63SPOT4        # BRANCH IF ANTENNA ALREADY IN POSITION 1

                CAF     CODE500         # ASTRONAUT:    PLEASE CRANK THE
                TC      BANKCALL        #               SILLY THING AROUND
                CADR    GOPERF1
                TCF     GOTOPOOH        # TERMINATE
                TCF     P63SPOT3        # PROCEED       SEE IF HE'S LYING

P63SPOT4        TC      BANKCALL        # ENTER         INITIALIZE LANDING RADAR
                CADR    SETPOS1

                TC      POSTJUMP        # OFF TO SEE THE WIZARD ...
                CADR    BURNBABY
```

Quake III source code (excerpt)

```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    x2 = number * 0.5F;
    y = number;
    i = * ( long * ) &y; // evil floating point bit level hacking
    i = 0x5f3759df - ( i >> 1 ); // what the fuck?
    y = * ( float * ) &i;
    y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
// y = y * ( threehalfs - ( x2 * y * y ) ); // 2nd iteration, this
can be removed

    return y;
}
```

Software *source code* is precious knowledge

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

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

Apollo 11 source code (excerpt)

```
P63SPOT3        CA      BIT6            # IS THE LR ANTENNA IN POSITION 1 YET
                EXTEND
                RAND    CHAN33
                EXTEND
                BZF     P63SPOT4        # BRANCH IF ANTENNA ALREADY IN POSITION 1

                CAF     CODE500         # ASTRONAUT:    PLEASE CRANK THE
                TC      BANKCALL        #               SILLY THING AROUND
                CADR    GOPERF1
                TCF     GOTOPOOH        # TERMINATE
                TCF     P63SPOT3        # PROCEED       SEE IF HE'S LYING

P63SPOT4        TC      BANKCALL        # ENTER         INITIALIZE LANDING RADAR
                CADR    SETPOS1

                TC      POSTJUMP        # OFF TO SEE THE WIZARD ...
                CADR    BURNBABY
```

Quake III source code (excerpt)

```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    x2 = number * 0.5F;
    y = number;
    i = * ( long * ) &y; // evil floating point bit level hacking
    i = 0x5f3759df - ( i >> 1 ); // what the fuck?
    y = * ( float * ) &i;
    y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
// y = y * ( threehalfs - ( x2 * y * y ) ); // 2nd iteration, this
can be removed

    return y;
}
```

Len Shustek, Computer History Museum 2006

"Source code provides a view into the mind of the designer."

Calling for source code preservation: UNESCO

Experts call for greater recognition of software source code as heritage for sustainable development

.6 November 2018

UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …

Calling for source code preservation: UNESCO

Experts call for greater recognition of software source code as heritage for sustainable development

.6 November 2018

UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …

The call is published on Feb 2019

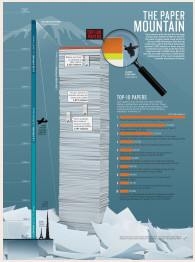
Calling for source code preservation: UNESCO

UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …

The call is published on Feb 2019

"[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive"
`https://en.unesco.org/foss/paris-call-software-source-code`

Software powers modern research

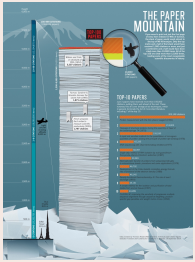


[…] software […] essential in their fields.
Top 100 papers (Nature, 2014)

Sometimes, if you dont have the software, you dont have the data
Christine Borgman, Paris, 2018

Source code history — for open science

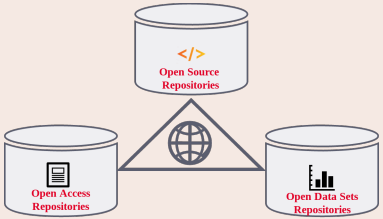
Software powers modern research



[…] software […] essential in their fields.
Top 100 papers (Nature, 2014)

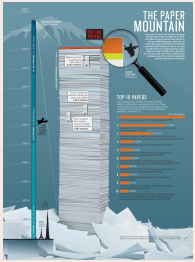
Sometimes, if you dont have the software, you dont have the data
Christine Borgman, Paris, 2018

Missing pillar: software (source code)



Source code history — for open science

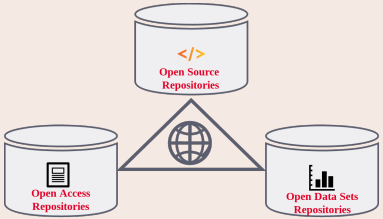
Software powers modern research



[…] software […] essential in their fields.
Top 100 papers (Nature, 2014)

Sometimes, if you dont have the software, you dont have the data
Christine Borgman, Paris, 2018

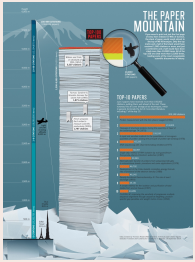
Missing pillar: software (source code)



The links in the picture are important

Source code history — for open science

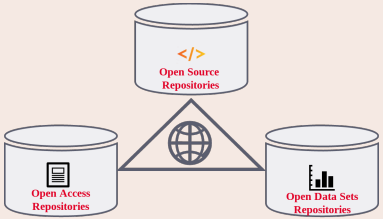
Software powers modern research



[…] software […] essential in their fields.
Top 100 papers (Nature, 2014)

Sometimes, if you dont have the software, you dont have the data
Christine Borgman, Paris, 2018

Missing pillar: software (source code)



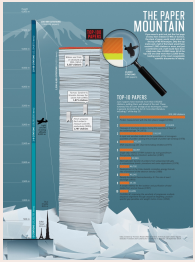
The links in the picture are important

Nota Bene

software may be a *tool*, a *research outcome* and a *research objet*

Source code history — for open science

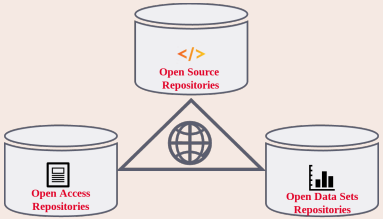
Software powers modern research



[…] software […] essential in their fields.
Top 100 papers (Nature, 2014)

Sometimes, if you dont have the software, you dont have the data
Christine Borgman, Paris, 2018

Missing pillar: software (source code)



The links in the picture are important

Nota Bene

software may be a *tool*, a *research outcome* and a *research objet*

access to the *source code* is essential!

Preserving the history of source code is important for *reproducibility*

Source code history — for security and transparency

Where does reused software come from?

Source code history — for security and transparency

Where does reused software come from?

Do *you* know where it comes from?

- the software you ship
- the software you use
- the software you acquire
- the software that
  - has that bug
  - has that vulnerability

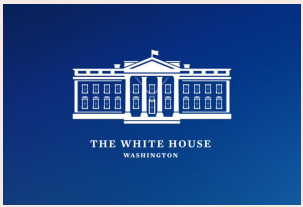
Source code history — for security and transparency

Where does reused software come from?

Do *you* know where it comes from?

- the software you ship
- the software you use
- the software you acquire
- the software that
  - has that bug
  - has that vulnerability

KYSW: Know Your SoftWare



Like KYC in banking, KYSW is now essential all over IT…

Sec. 4. Enhancing Software Supply Chain Security

ensuring and attesting, to the extent practicable, to the integrity and provenance of open source software

May 2021 POTUS Executive Order

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

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 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

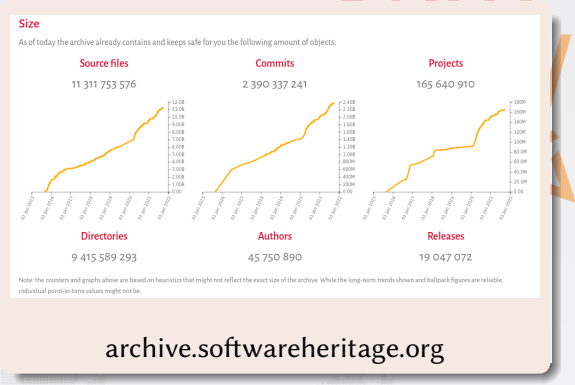
preserve all software source code

Research infrastructure

enable analysis of all software source code

The largest public source code archive, principled

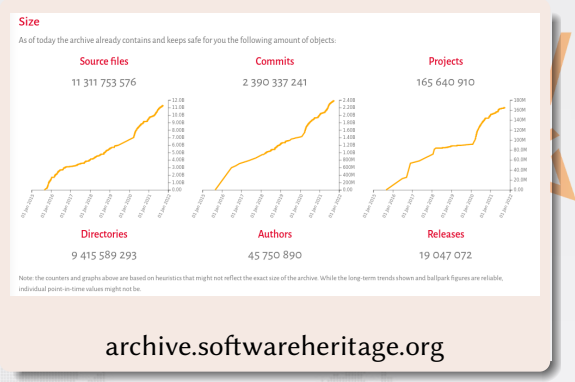
bit.ly/swhpaper



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

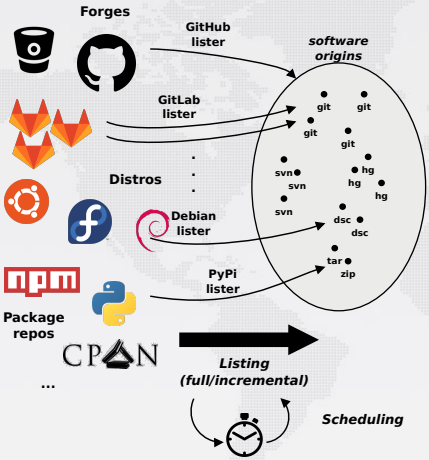
Content (billions!)

- intrinsic identifiers
- facts and provenance

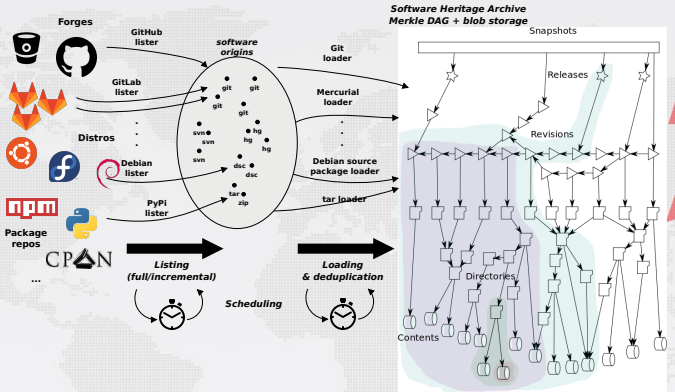
Organization

- non-profit
- multi-stakeholder

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

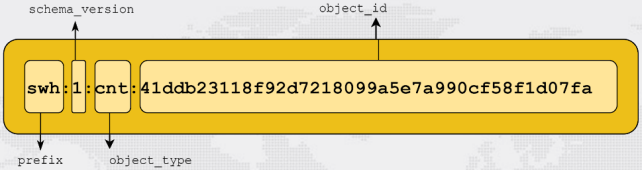


A peek under the hood: a global view on the software commons

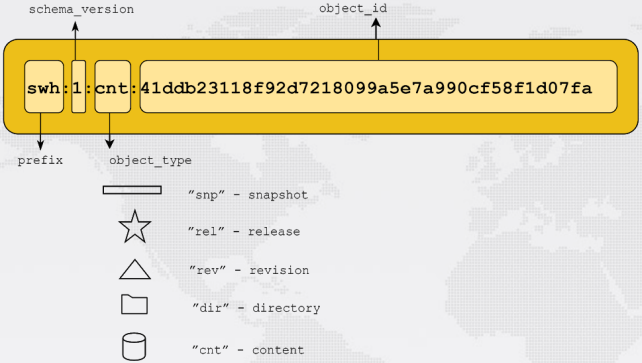


A global graph linking together fully deduplicated source code artifact (files, commits, directories, releases, etc.) to the places that distribute them (e.g., Git repositories), providing a unified view on the entire *Software Commons*.
Size: ~20 B nodes, ~200 B edges, ~600 TB (uncompressed) blobs

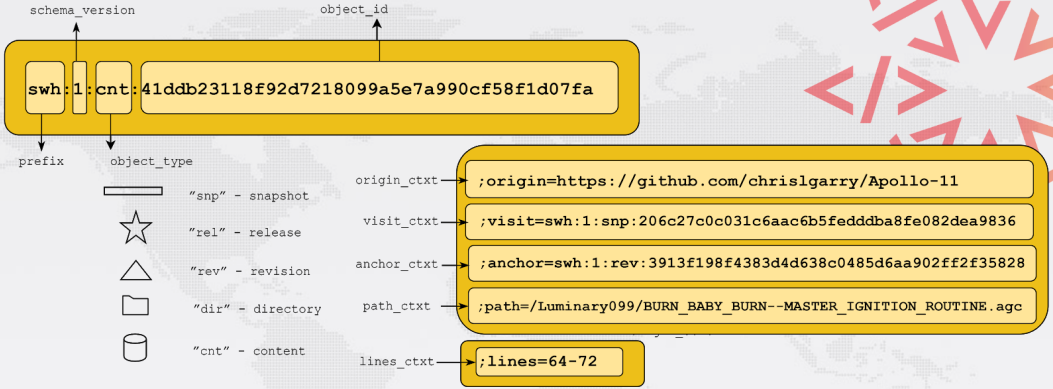
Software Heritage *intrinsic* Identifiers (SWHID) (full spec)



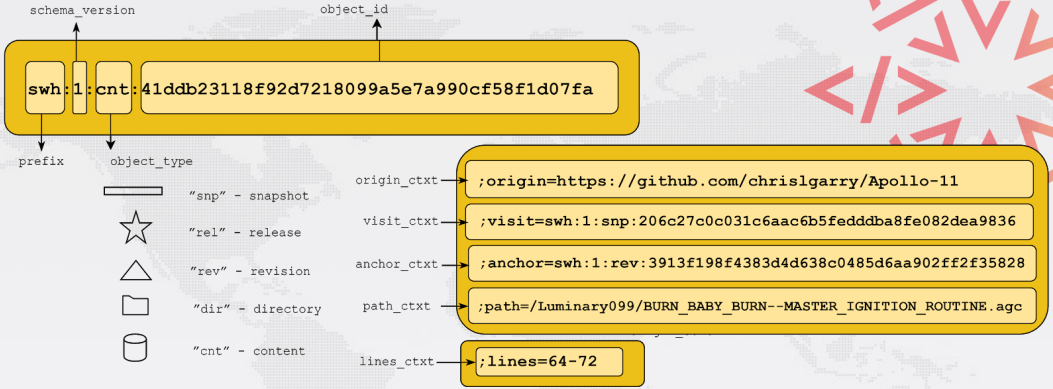
Software Heritage *intrinsic* Identifiers (SWHID) (full spec)



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



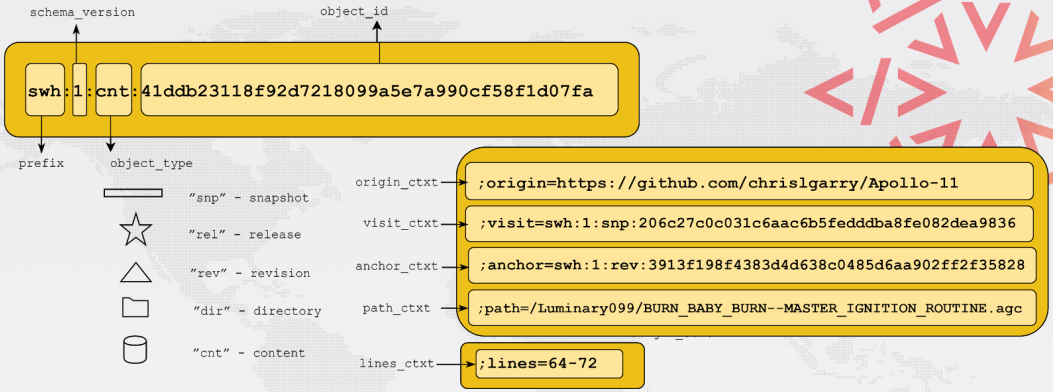
Software Heritage *intrinsic* Identifiers (SWHID) (full spec)



An emerging standard

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

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



An emerging standard

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

Examples:

- Apollo 11 AGC excerpt
- Quake III rsqrt

Outline

A walkthrough

- Browse the archive
- Trigger archival of your preferred software in a breeze
- Get and use SWHIDs (full spec available online)
- The Apollo 11 AGC source code example
- Cite software with the biblatex-software style from CTAN
- Example use in a research article: compare Fig. 1 and conclusions
  - in the 2012 version
  - in the updated version using SWHIDs and Software Heritage
- Example in a journal: an article from IPOL
- Curated deposit in SWH via HAL, see for example: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, …
- Rescue landmark legacy software, see the SWHAP process with UNESCO

Outline

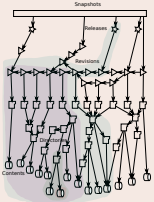
A revolutionary research infrastructure designed for source code

The *graph* of Software Development

All software development in a single graph …

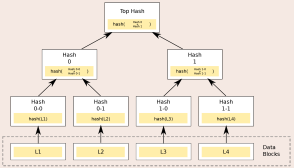
A revolutionary research infrastructure designed for source code

The *graph* of Software Development



All software development in a single graph …

The *blockchain* of Software Development



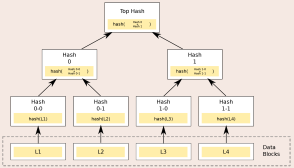
… a single Merkle graph!

A revolutionary research infrastructure designed for source code

The *graph* of Software Development

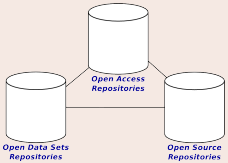
All software development in a single graph …

The *blockchain* of Software Development



… a single Merkle graph!

A *pillar* of Open Science



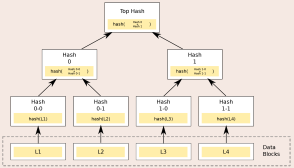
Reference archive of Research Software

A revolutionary research infrastructure designed for source code

The *graph* of Software Development

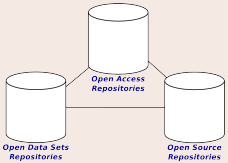
All software development in a single graph …

The *blockchain* of Software Development



… a single Merkle graph!

A *pillar* of Open Science



Reference archive of Research Software

Reference platform for *Big Code*

A single, uniform data structure

A challenging scientific and technical undertaking

A novel, large infrastructure

- gigantic Merkle graph
- object storage with peculiar workload
- simple problems become hard, e.g., counting tens of billions of objects, or sorting all possible origins of a node

A challenging scientific and technical undertaking

A novel, large infrastructure

- gigantic Merkle graph
- object storage with peculiar workload
- simple problems become hard, e.g., counting tens of billions of objects, or sorting all possible origins of a node

First dataset available as open data

Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli
The Software Heritage Graph Dataset: Public software development under one roof
MSR 2019: 16th Intl. Conf. on Mining Software Repositories. IEEE
preprint: `http://deb.li/swhmsr19`

- used as topic for the MSR 2020 mining competition

A challenging scientific and technical undertaking

A novel, large infrastructure

- gigantic Merkle graph
- object storage with peculiar workload
- simple problems become hard, e.g., counting tens of billions of objects, or sorting all possible origins of a node

First dataset available as open data

Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli
The Software Heritage Graph Dataset: Public software development under one roof
MSR 2019: 16th Intl. Conf. on Mining Software Repositories. IEEE
preprint: `http://deb.li/swhmsr19`

- used as topic for the MSR 2020 mining competition

→ for more, see last week's talk with the DIG team: Analyzing the Global Graph of Public Software Development upsilon.cc/~zack/talks/2021/2021-10-14-telecom-dig.pdf

Outline

Focus on Academia: growing adoption (selection)

HAL software curated deposit workflow

Curated Archiving of Research Software Artifacts
International Journal of Digical Curation, 2020

Reference archive for swmath.org

See *code* links, e.g. SemiPar package

Focus on Academia: growing adoption (selection)

HAL software curated deposit workflow

Curated Archiving of Research Software Artifacts
International Journal of Digical Curation, 2020

Reference archive for swmath.org

swMATH
an information service for mathematical software

See *code* links, e.g. SemiPar package

IPOL (image processing)

- archive (deposit)
- reference
- BibLaTeX

eLife (life sciences)

- archive (save code now)
- reference

JTCAM (mechanics)

- instructions for authors
- biblatex-software in journal LaTeX class

Focus on Academia: growing adoption (selection)

HAL software curated deposit workflow

Curated Archiving of Research Software Artifacts
International Journal of Digical Curation, 2020

Reference archive for swmath.org

swMATH
an information service for mathematical software

See *code* links, e.g. SemiPar package

IPOL (image processing)

- archive (deposit)
- reference
- BibLaTeX

eLife (life sciences)

- archive (save code now)
- reference

JTCAM (mechanics)

- instructions for authors
- biblatex-software in journal LaTeX class

Policy: France

NATIONAL PLAN FOR OPEN SCIENCE

National Plan for Open Science

Policy: Europe

EOSC SIRS report

- SWHIDs
- archive

Guidelines

- summary
- ICMS 2020

Sharing the vision

UNESCO
United Nations
Educational, Scientific and
Cultural Organization

And many more ...
www.softwareheritage.org/support/testimonials

Sharing the vision

UNESCO
United Nations Educational, Scientific and Cultural Organization

And many more ...
www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors

You may help!

Foster the adoption of research best practices

- archive and reference relevant source code (save code now, and deposit)
- use Software Heritage and biblatex-software in articles, journals, and books
- rescue and preserve landmark legacy source code with SWHAP

You may help!

Foster the adoption of research best practices

- archive and reference relevant source code (save code now, and deposit)
- use Software Heritage and biblatex-software in articles, journals, and books
- rescue and preserve landmark legacy source code with SWHAP

Engage with Software Heritage as a researcher

- use the archive for your own software-related experiments
- work with us to tackle open technical and research problems

You may help!

Foster the adoption of research best practices

- archive and reference relevant source code (save code now, and deposit)
- use Software Heritage and biblatex-software in articles, journals, and books
- rescue and preserve landmark legacy source code with SWHAP

Engage with Software Heritage as a researcher

- use the archive for your own software-related experiments
- work with us to tackle open technical and research problems

Engage with Software Heritage as an organization

- become a member/sponsor
- build a Software Heritage mirror
- contribute to the preservation mission

Thank you!

Resources

archive	archive.softwareheritage.org
stay posted	softwareheritage.org/newsletter
blog	softwareheritage.org/blog

References (selected; full list at softwareheritage.org/publications)

Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchiroli
Building the Universal Archive of Source Code
Communication of the ACM, October 2018

Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli
The Software Heritage Graph Dataset: Large-scale Analysis of Public Software Development History
MSR 2020: 17th Intl. Conf. on Mining Software Repositories. IEEE

Roberto Di Cosmo
Archiving and Referencing Source Code with Software Heritage
International Congress on Mathematical Software (ICMS), 2020

MESRI
Second French Plan for Open Science
www.ouvrirlascience.fr/second-national-plan-for-open-science, 2001

Archiving goals

Targets: VCS repositories & source code releases (e.g., tarballs, packages)

We DO archive

- file content (= blobs)
- revisions (= commits), with full metadata
- releases (= tags), ditto
- where (origin) & when (visit) we found any of the above

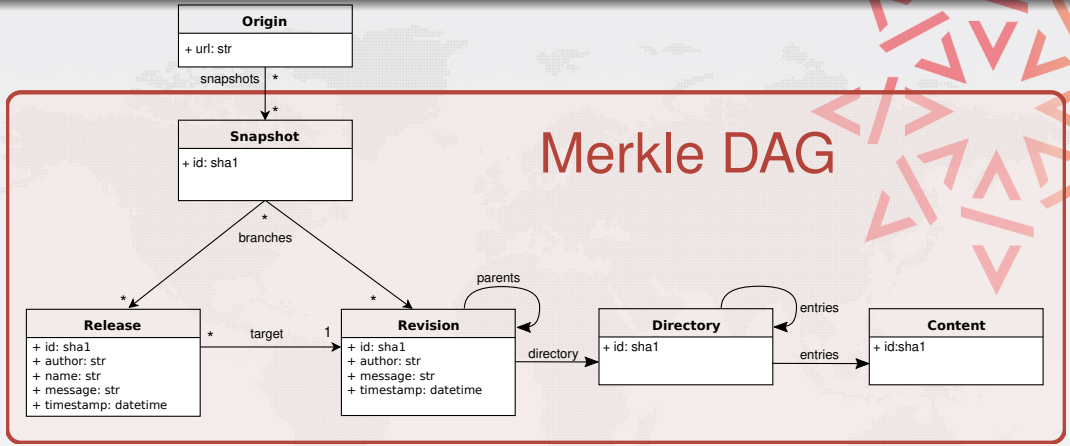
… in a VCS-/archive-agnostic canonical data model

We DON'T archive (yet)

- homepages, wikis
- BTS/issues/code reviews/etc.
- mailing lists

Long term vision: play our part in a *"semantic wikipedia of software"*

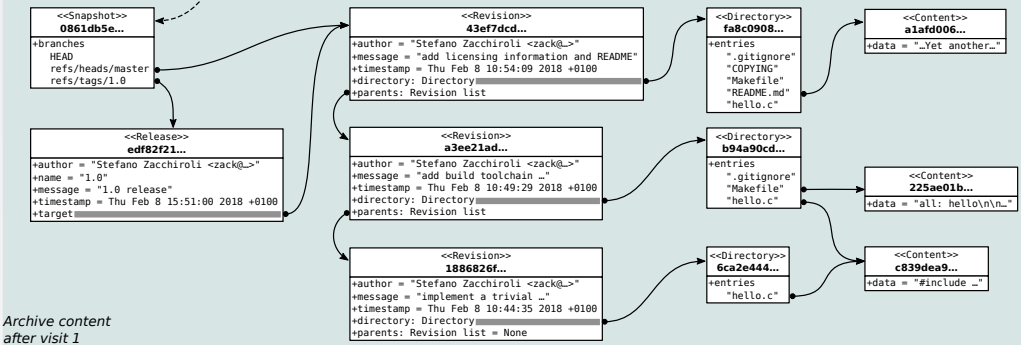
Data model



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*.

The archive: a (giant) Merkle DAG

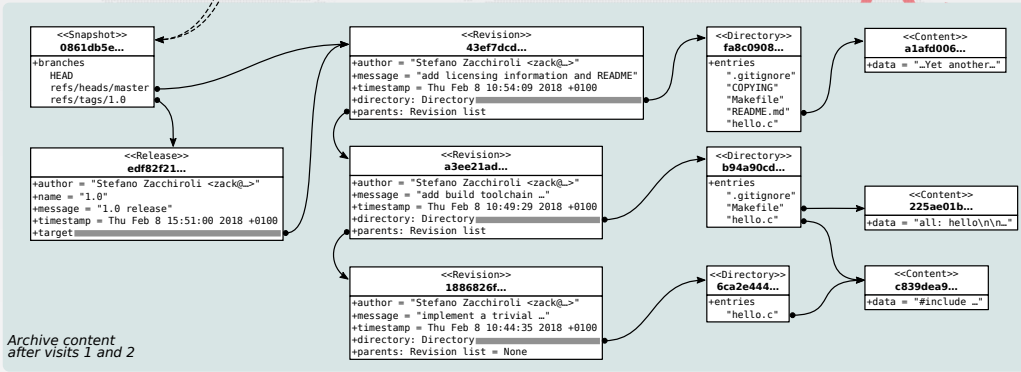
origin	visit	snapshot	timestamp
https://forge.softwareheritage.org/source/helloworld.git	1	0861db5e…	Fri Feb 9 12:38:45 2018 +0100



Archive content after visit 1

The archive: a (giant) Merkle DAG

origin	visit	snapshot	timestamp
https://forge.softwareheritage.org/source/helloworld.git	1	0861db5e...	Fri Feb 9 12:38:45 2018 +0100
https://forge.softwareheritage.org/source/helloworld.git	2	0861db5e...	Fri Feb 9 13:29:00 2018 +0100



The archive: a (giant) Merkle DAG