

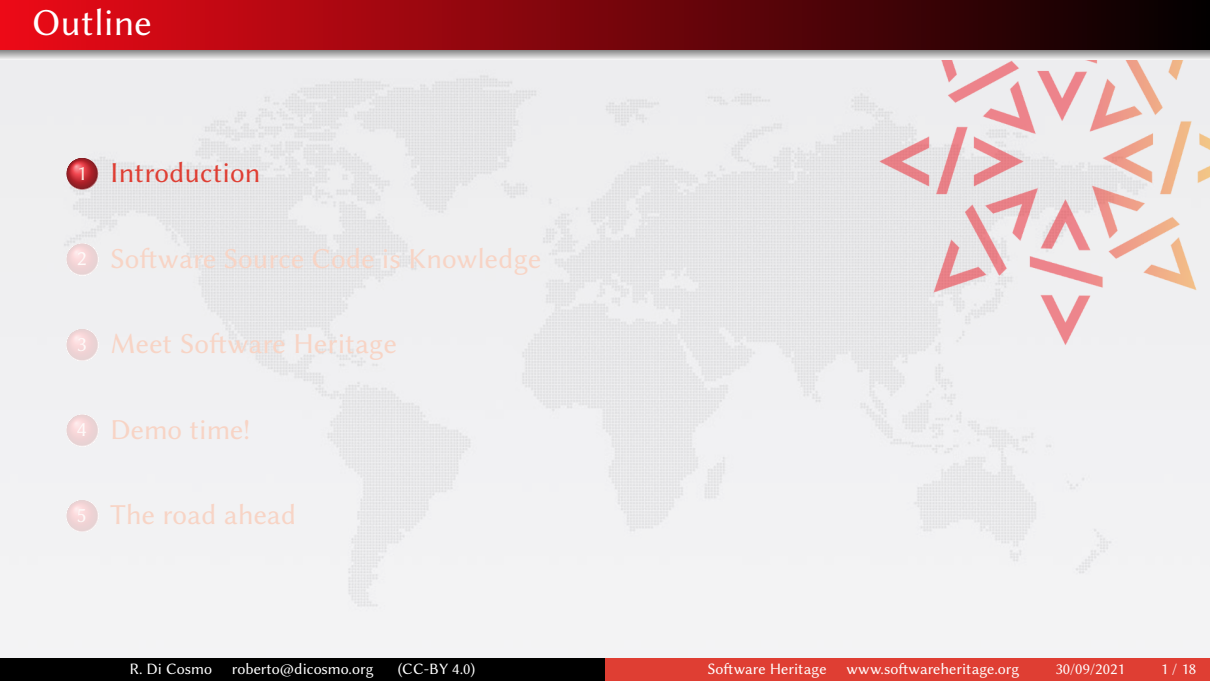
Archiver, référencer et citer les codes source de la recherche

Roberto Di Cosmo
Inria and Université de Paris

30/09/2021



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

- 
- 1 Introduction
 - 2 Software Source Code is Knowledge
 - 3 Meet Software Heritage
 - 4 Demo time!
 - 5 The road ahead

Computer Science professor in Paris, now working at INRIA

- 30 years of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- 20 years of Free and Open Source Software
- 10 years building and directing structures for the common good



1999 *DemoLinux* – first live GNU/Linux distro

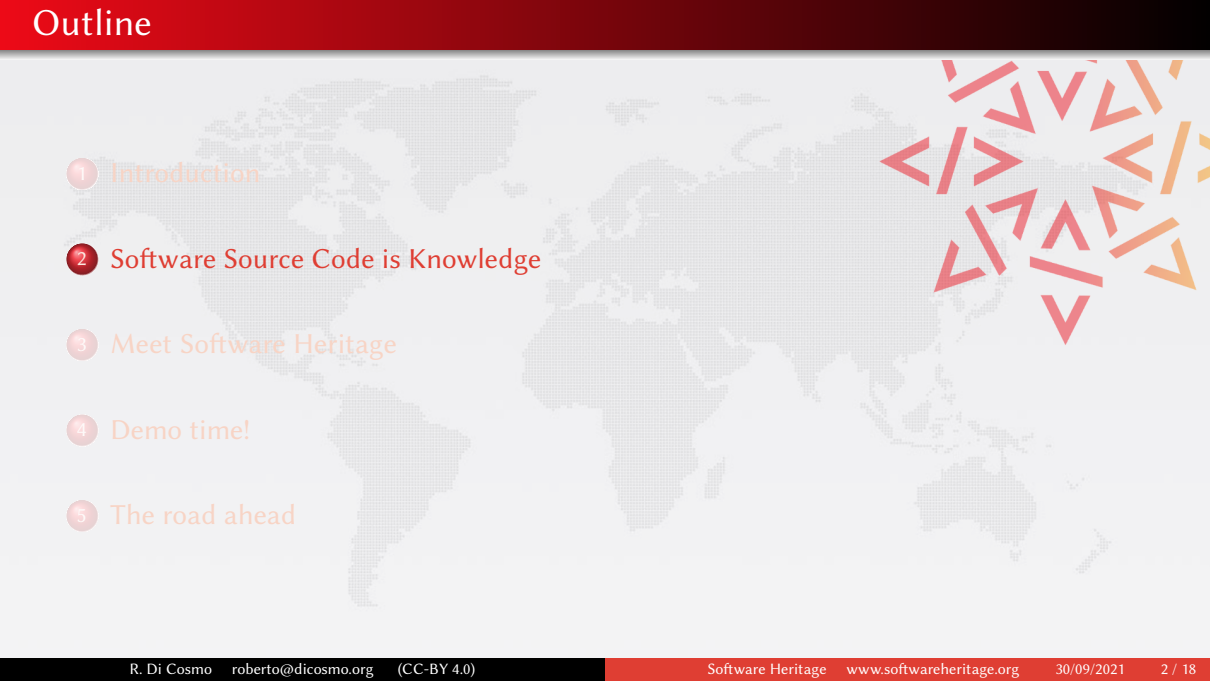
2007 *Free Software Thematic Group*
150 members 40 projects 200Me

2008 *Mancoosi project* www.mancoosi.org

2010 *IRILL* www.irill.org

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

- 
- 1 Introduction
 - 2 Software Source Code is Knowledge
 - 3 Meet Software Heritage
 - 4 Demo time!
 - 5 The road ahead

Why Software *Source Code*

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.”

Why Software Source Code

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)

```
P63SP0T3      CA      BIT6          # IS THE LR ANTENNA IN POSITION 1 YET
               EXTEND
               RAND      CHAN33
               EXTEND
               BZF      P63SP0T4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

               CAF      CODE500      # ASTRONAUT:  PLEASE CRANK THE
               TC       BANKCALL      #              SILLY THING AROUND
               CADR      GOPERF1
               TCF      GOTOP00H      # TERMINATE
               TCF      P63SP0T3      # PROCEED      SEE IF HE'S LYING

P63SP0T4      TC       BANKCALL      # ENTER      INITIALIZE LANDING RADAR
               CADR      SETPOS1

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

Why Software Source Code

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)

```
P63SP0T3      CA      BIT6      # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND      CHAN33
              EXTEND
              BZF      P63SP0T4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF      CODE500      # ASTRONAUT: PLEASE CRANK THE
              TC      BANKCALL      # SILLY THING AROUND
              CADR      GOPERF1
              TCF      GOTOP00H      # TERMINATE
              TCF      P63SP0T3      # PROCEED SEE IF HE'S LYING

P63SP0T4      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;
}
```

Why Software Source Code

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)

```
P63SP0T3      CA      BIT6      # IS THE LR ANTENNA IN POSITION 1 YET
              EXTEND
              RAND      CHAN33
              EXTEND
              BZF      P63SP0T4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

              CAF      CODE500      # ASTRONAUT: PLEASE CRANK THE
              TC      BANKCALL      # SILLY THING AROUND
              CADR      GOPERF1
              TCF      GOTOP00H      # TERMINATE
              TCF      P63SP0T3      # PROCEED SEE IF HE'S LYING

P63SP0T4      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.”

Source code is *special* (software is *not* data)

Software *evolves* over time

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

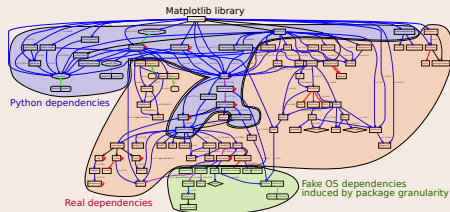
Source code is *special* (software is *not* data)

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
 - *research software* a thin top layer
- sophisticated *developer communities*



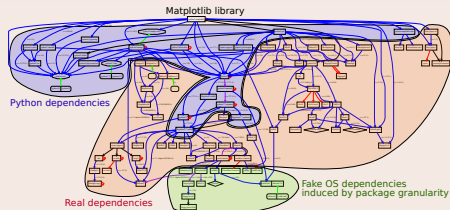
Source code is *special* (software is *not* data)

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
 - *research software* a thin top layer
- sophisticated *developer communities*



Precious, endangered *executable* and *human readable* knowledge

key people **passing away**, platforms (GoogleCode, Gitorious, etc.) closing down ...

no organised effort to catalog and archive it

Source code is *special*, cont'd

Software is complex

- Structure** monolithic/composite; self-contained/external dependencies
- Lifetime** one-shot/long term
- Community** one man/one team/distributed community
- Authorship** multiple roles: Architecture, Management, Development, Documentation, Testing, ...
- Authority** institutions/organizations/communities/single person

Source code is *special*, cont'd

Software is complex

- Structure** monolithic/composite; self-contained/external dependencies
- Lifetime** one-shot/long term
- Community** one man/one team/distributed community
- Authorship** multiple roles: Architecture, Management, Development, Documentation, Testing, ...
- Authority** institutions/organizations/communities/single person

Versioning, granularity

- Project** “Inria created OCaml and Scikit-learn”

Source code is *special*, cont'd

Software is complex

- Structure** monolithic/composite; self-contained/external dependencies
- Lifetime** one-shot/long term
- Community** one man/one team/distributed community
- Authorship** multiple roles: Architecture, Management, Development, Documentation, Testing, ...
- Authority** institutions/organizations/communities/single person

Versioning, granularity

- Project** “Inria created OCaml and Scikit-learn”
- Release** “2D Voronoi Diagrams were introduced in CGAL 3.1.0”

Source code is *special*, cont'd

Software is complex

- Structure** monolithic/composite; self-contained/external dependencies
- Lifetime** one-shot/long term
- Community** one man/one team/distributed community
- Authorship** multiple roles: Architecture, Management, Development, Documentation, Testing, ...
- Authority** institutions/organizations/communities/single person

Versioning, granularity

- Project** “Inria created OCaml and Scikit-learn”
- Release** “2D Voronoi Diagrams were introduced in CGAL 3.1.0”
- Exact state of a project** “This result was produced using commit 0064fbd...”

Source code is *special*, cont'd

Software is complex

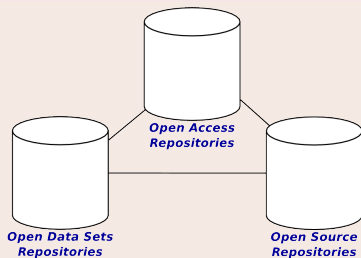
- Structure** monolithic/composite; self-contained/external dependencies
- Lifetime** one-shot/long term
- Community** one man/one team/distributed community
- Authorship** multiple roles: Architecture, Management, Development, Documentation, Testing, ...
- Authority** institutions/organizations/communities/single person

Versioning, granularity

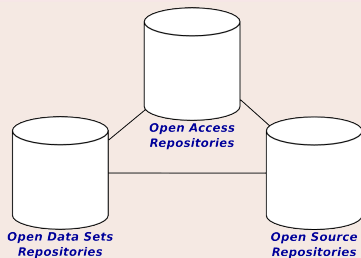
- Project** “Inria created OCaml and Scikit-learn”
- Release** “2D Voronoi Diagrams were introduced in CGAL 3.1.0”
- Exact state of a project** “This result was produced using commit 0064fbd...”
- Code fragment** “The core algorithm is in lines 101 to 143 of the file parmap.ml contained in the precise state of the project corresponding to commit 0064fbd...”

Software Source code: a pillar of Open Science

Three pillars of Open Science



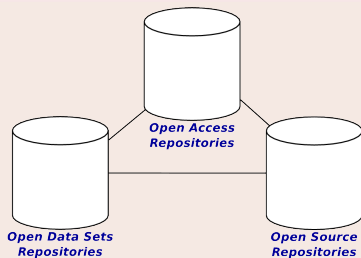
Three pillars of Open Science



A plurality of needs

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

Three pillars of Open Science



A plurality of needs

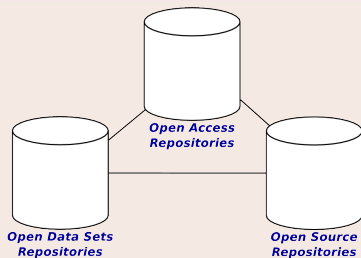
Researcher

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

Laboratory/team track software contributions

- produce reports / web page

Three pillars of Open Science



A plurality of needs

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

- Laboratory/team** track software contributions
- produce reports / web page

- Research Organization** know its **software assets**
- technology **transfer**
 - impact **metrics**

Archive

Research software artifacts must be properly **archived**
make sure we can *retrieve* them (*reproducibility*)

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*)

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*)

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!*)

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!*)

We need an infrastructure *designed for* software source code:

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*)

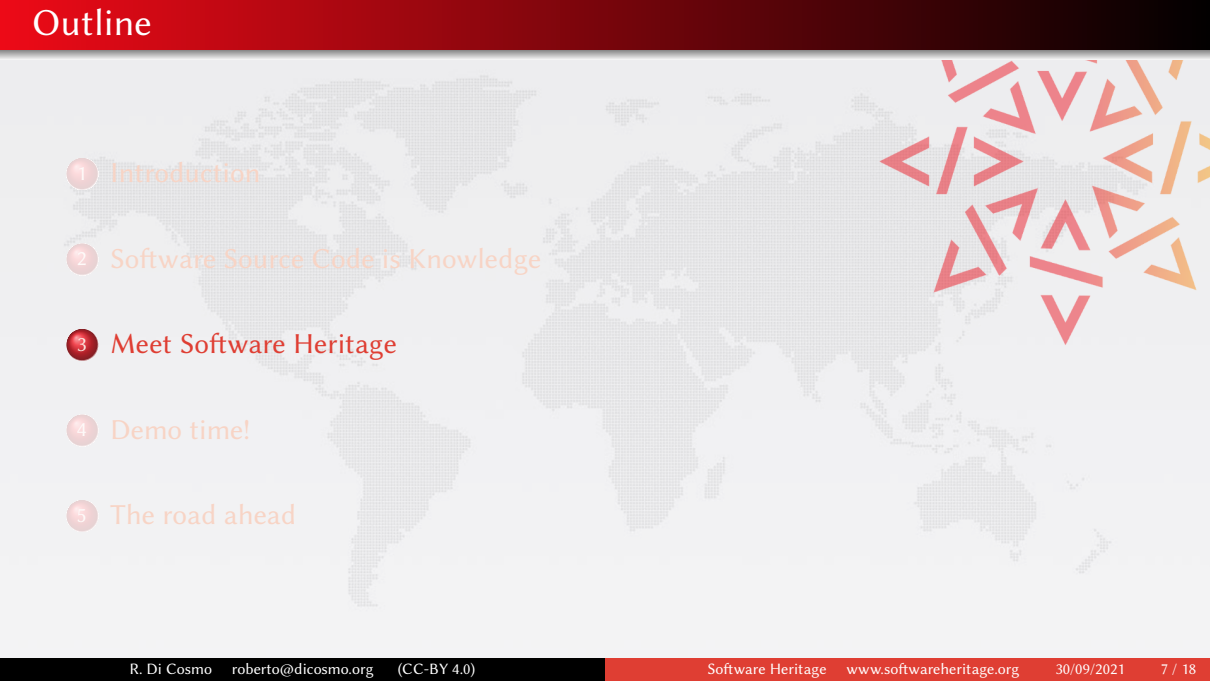
Describe

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!*)

We need an infrastructure *designed for* software source code: *now we have one!*

- 
- 1 Introduction
 - 2 Software Source Code is Knowledge
 - 3 Meet Software Heritage**
 - 4 Demo time!
 - 5 The road ahead



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



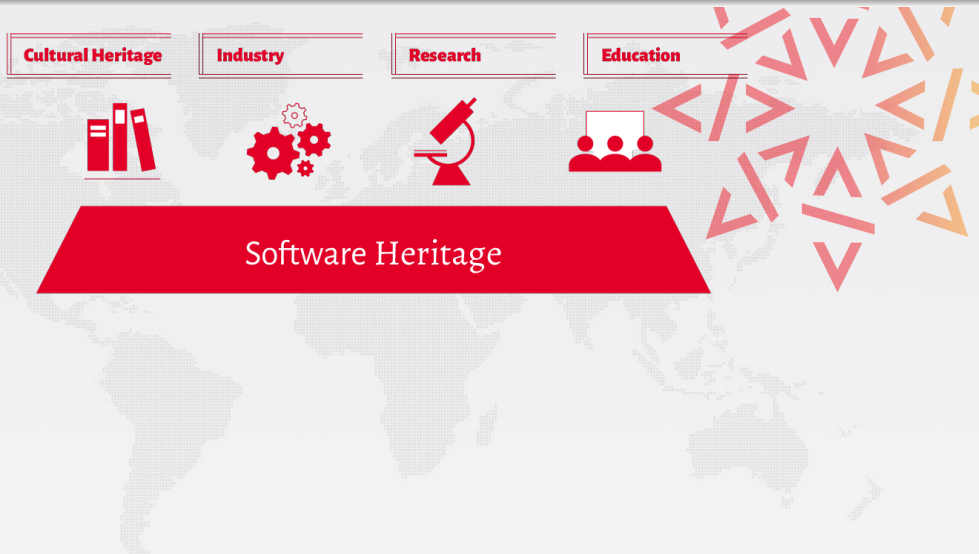
preserve all software
source code

Research infrastructure



enable analysis of all
software source code

The largest software archive, a shared infrastructure



The largest software archive, a shared infrastructure

Cultural Heritage



Industry



Research



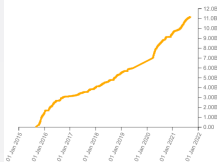
Education



Software Heritage

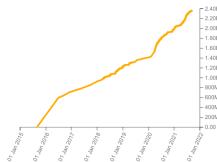
Source files

11,218,541,481



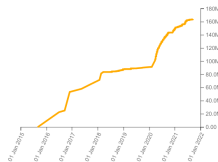
Commits

2,368,202,867



Projects

164,455,982



Directories

9,363,834,851

Authors

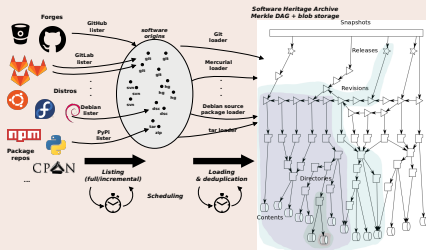
45,586,580

Releases

18,891,129

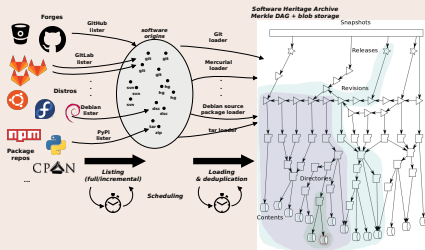
Addressing the four needs (see ICMS 2020 for details)

Archive (10B+ files, 150M+ projects)



Addressing the four needs (see ICMS 2020 for details)

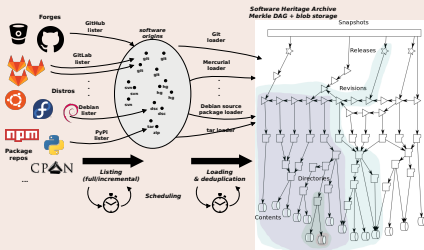
Archive (10B+ files, 150M+ projects)



- save.softwareheritage.org
- deposit.softwareheritage.org

Addressing the four needs (see ICMS 2020 for details)

Archive (10B+ files, 150M+ projects)



- save.softwareheritage.org
- deposit.softwareheritage.org

Reference (20 billion SWHIDs)

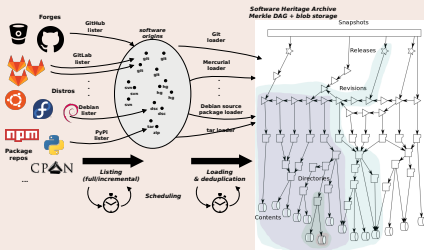
Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs



Now supported in SPDX 2.2, Wikidata etc.

Addressing the four needs (see ICMS 2020 for details)

Archive (10B+ files, 150M+ projects)



- save.softwareheritage.org
- deposit.softwareheritage.org

Describe

- *Intrinsic metadata* from source code
- Contributed the **Codemeta** generator

Reference (20 billion SWHIDs)

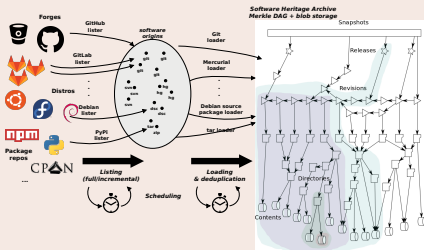
Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs



Now supported in SPDX 2.2, Wikidata etc.

Addressing the four needs (see ICMS 2020 for details)

Archive (10B+ files, 150M+ projects)



- save.softwareheritage.org
- deposit.softwareheritage.org

Describe

- *Intrinsic metadata* from source code
- Contributed the **Codemeta** generator

Reference (20 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs



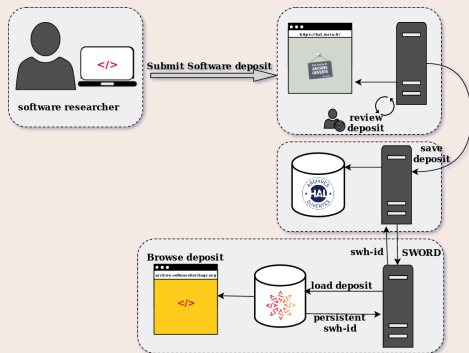
Now supported in SPDX 2.2, Wikidata etc.

Cite/Credit

- Contributed *software citation* style
biblatex-software, v 1.2-2 now on CTAN

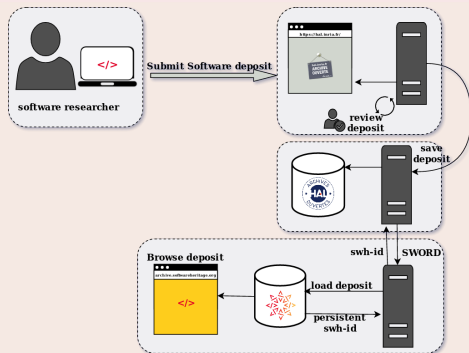
Deposit software in HAL

<http://hal.inria.fr/hal-01738741>



Deposit software in HAL

<http://hal.inria.fr/hal-01738741>

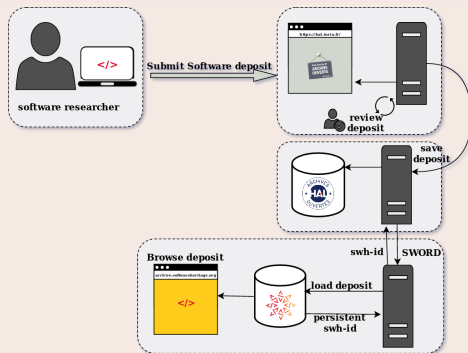


Generic mechanism:

- SWORD based
- review process
- versioning

Deposit software in HAL

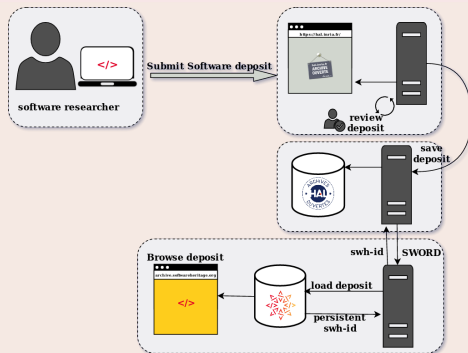
<http://hal.inria.fr/hal-01738741>



Generic mechanism:

- SWORD based
- review process
- versioning
- **today**: deposit .zip or .tar.gz file (*guide*)
- **under development**:
 - just provide *SWHID*, and metadata will be retrieved if present

Deposit software in HAL

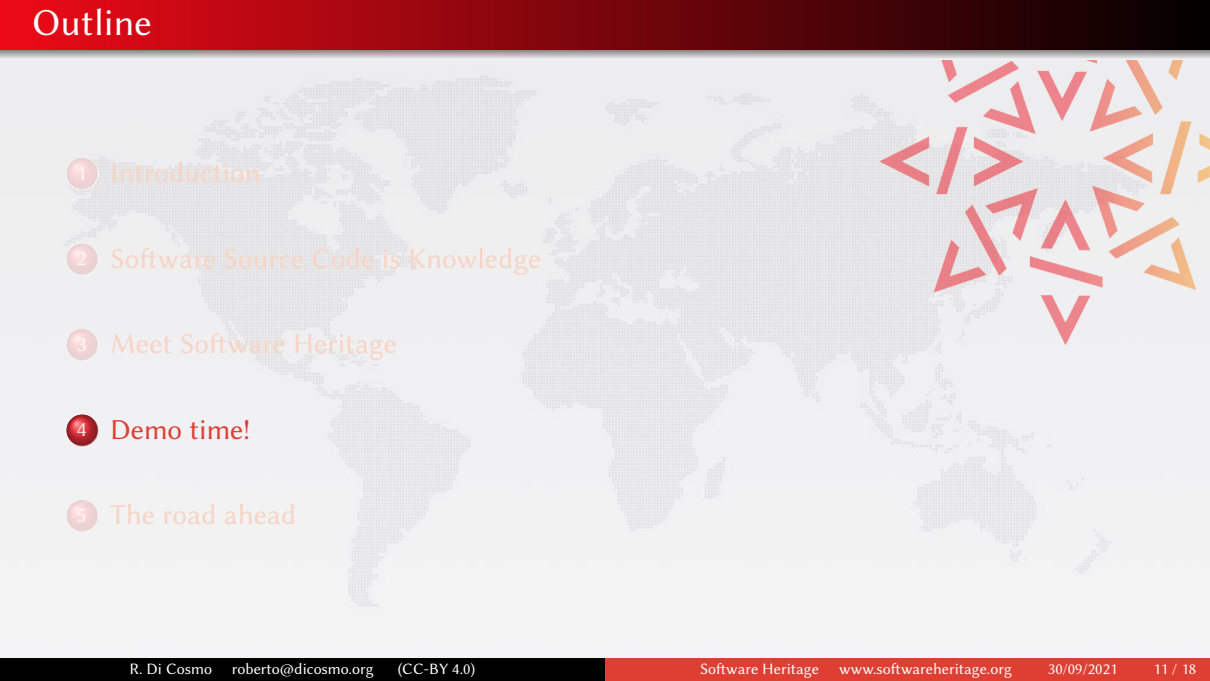
<http://hal.inria.fr/hal-01738741>

Generic mechanism:

- SWORD based
- review process
- versioning
- **today:** deposit .zip or .tar.gz file (*guide*)
- **under development:**
 - just provide *SWHID*, and metadata will be retrieved if present

Deposit/describe research software in HAL

- author: <https://hal.archives-ouvertes.fr/hal-01872189>
- moderator: <https://hal.archives-ouvertes.fr/hal-01876705>

- 
- 1 Introduction
 - 2 Software Source Code is Knowledge
 - 3 Meet Software Heritage
 - 4 Demo time!**
 - 5 The road ahead

- Browse [the archive](#)
- [Trigger archival](#) of your preferred software in a breeze
- Get and use SWHIDs ([full specification available online](#))
- Example HAL deposits: [LinBox](#), [SLALOM](#), [Givaro](#), [NS2DDV](#), [SumGra](#), [Coq proof](#), ...
- Cite software [with the biblatex-software style](#) from CTAN
- Example in a real journal: [an article from IPOL](#)
- 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 use in a research article: extensive use of SWHIDs in [a replication experiment](#)
- Supporting reproducible builds: [Guix](#) and [Nix](#)

Summing up: a revolutionary infrastructure *designed for source code*



Software Heritage

www.softwareheritage.org

global source code archive

Library of Alexandria of source code



- harvest *all* software, not just research software
- *on demand harvesting* and *curated deposit*

universal intrinsic identifiers

SWHID standard is independent of version control systems

uniform data model, *full graph* of development history

enables large scale, big code research

Summing up: a revolutionary infrastructure *designed for source code*



Software Heritage

www.softwareheritage.org

global source code archive

Library of Alexandria of source code



- harvest *all* software, not just research software
- *on demand* harvesting and *curated deposit*

universal intrinsic identifiers

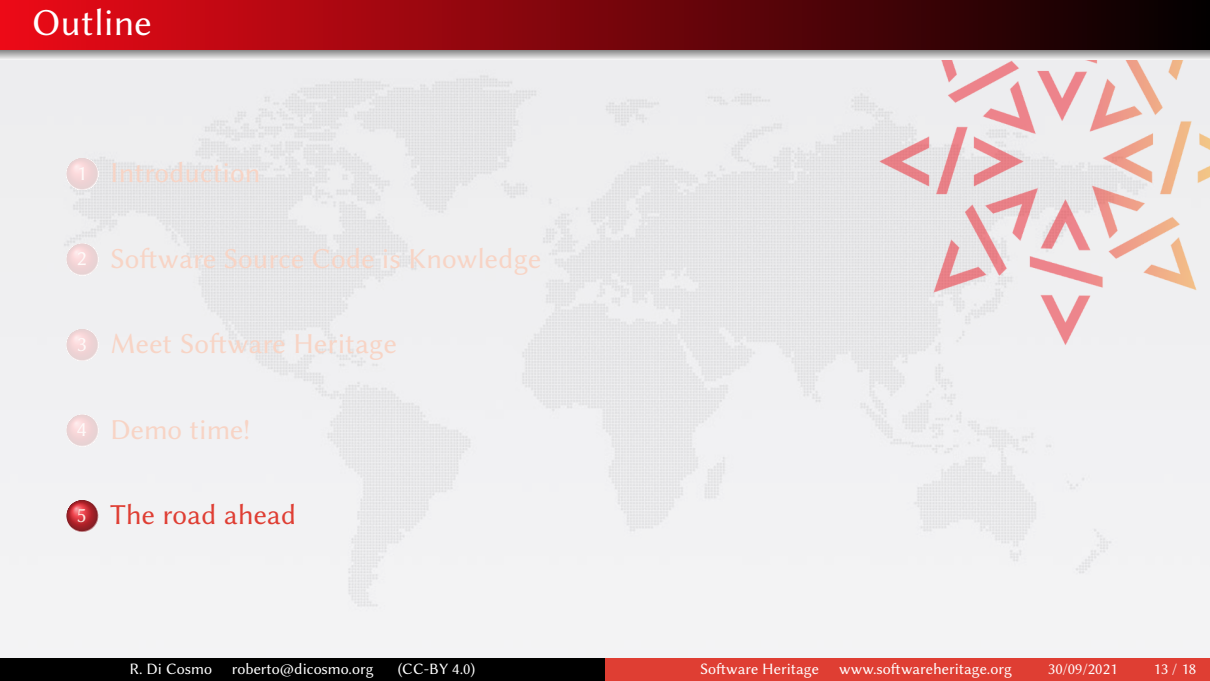
SWHID standard is independent of version control systems

uniform data model, *full graph* of development history

enables large scale, big code research

infrastructure for Open Science

base layer for software source code in the *Open Science architecture*

- 
- 1 Introduction
 - 2 Software Source Code is Knowledge
 - 3 Meet Software Heritage
 - 4 Demo time!
 - 5 The road ahead

Sharing the vision



United Nations
Educational, Scientific and
Cultural Organization



And many more ...

www.softwareheritage.org/support/testimonials

Sharing the vision



United Nations
Educational, Scientific and
Cultural Organization



And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors



Platinum sponsors



Gold sponsors



Silver sponsors



Bronze sponsors



Growing adoption in Academia (selection)

HAL software curated deposit workflow

Curated Archiving of Research Software Artifacts

International Journal of Digital Curation, 2020

Reference archive for swmath.org



See *code* links, e.g.

SemiPar package

Growing adoption in Academia (selection)

HAL software curated deposit workflow

Curated Archiving of Research Software Artifacts

International Journal of Digital Curation, 2020

Reference archive for swmath.org



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 $\text{\LaTeX}$ class

Growing adoption in Academia (selection)

HAL software curated deposit workflow

Curated Archiving of Research Software Artifacts

International Journal of Digital Curation, 2020

Reference archive for swmath.org



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 L^AT_EX class

Policy: France



*National Plan
Open Science*

Policy: Europe



EOSC SIRS report

- SWHIDs
- archive

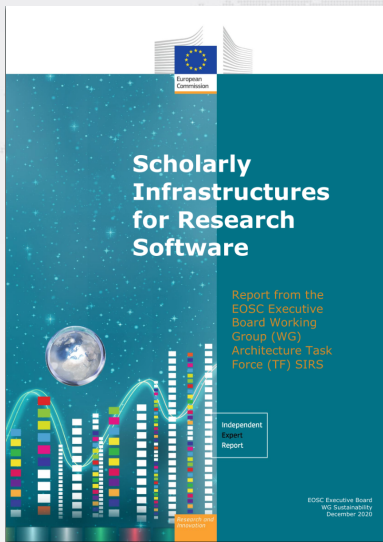
Guidelines

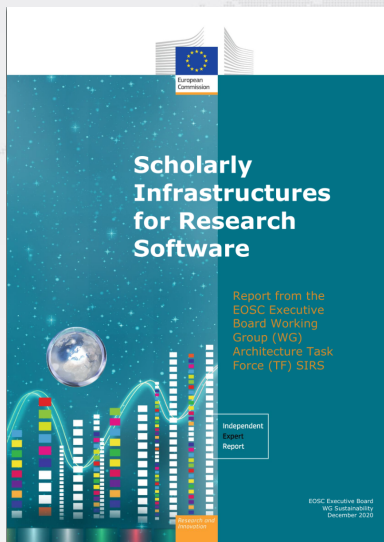


Software Heritage

- 1 Prepare your public repository
Readme, authors & license files
- 2 Save your code
<http://sw.here.org>
- 3 Reference your work
(Full repository, specific version or code fragments)

- summary
- ICMS 2020





Important *policy tool* in Open Science (Dec 2020)

- 9 infrastructures
 - 3 archives
 - 3 open access publishers
 - 3 aggregators
- recommendations
 - archive in Software Heritage, use SWHID
 - open non profit
 - default to open source for research software

"all research software should be made available under an Open Source license by default, and all deviations from this default practice should be properly motivated"

See <https://doi.org/10.2777/28598>

Focus: second French National Plan for Open Science



Focus: second French National Plan for Open Science



2nd National Plan for Open Science (6/7/2021)

Open and promote research software source code

- actions (selection)
 - charter for research software policy
 - recognize software development (see [announcement of the 2021 prize](#))
 - coordinate communities of practice
 - build a connected ecosystem of research outputs
- recommendations (selection)
 - archive in Software Heritage
 - standardise and use SWHID
 - build a national catalog of research software
 - leverage ADAC network

See [official announcement](#)

You may help!

Foster adoption and best practices for researchers

- describe and deposit via HAL
- archive and reference source code, cite using `bibtex-software`
- ask journals and conferences to use Software Heritage (like IPOL, etc.)

You may help!

Foster adoption and best practices for researchers

- describe and deposit via HAL
- archive and reference source code, cite using [biblatex-software](#)
- ask journals and conferences to use Software Heritage (like IPOL, etc.)

Engage with Software Heritage as an individual

- join the core team: *positions will be available*
- join the [ambassador program](#)
 - learn directly from the core team, share with your community

You may help!

Foster adoption and best practices for researchers

- describe and deposit via HAL
- archive and reference source code, cite using [biblatex-software](#)
- ask journals and conferences to use Software Heritage (like IPOL, etc.)

Engage with Software Heritage as an individual

- join the core team: *positions will be available*
- join the [ambassador program](#)
 - learn directly from the core team, share with your community

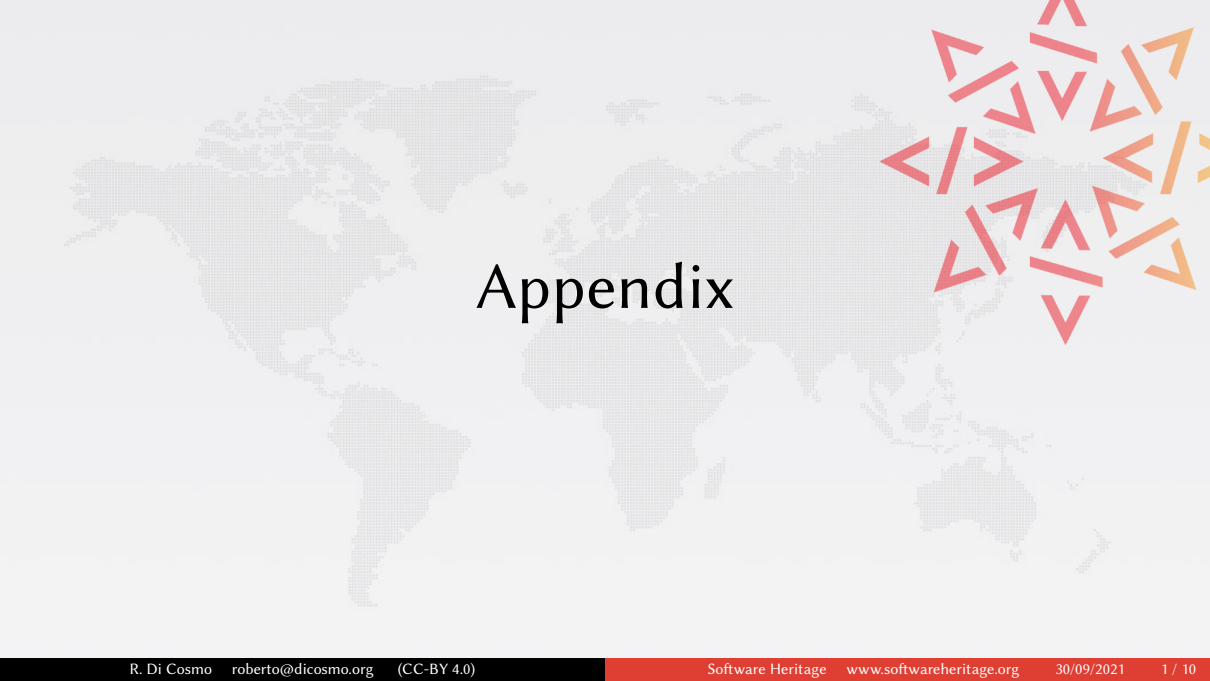
Engage with Software Heritage as an organization

- include software in your *Open Science policy*
- join as a [member/sponsor](#)
- establish a [mirror](#) (see e.g. ENEA): support the mission and enable Big Code

Questions?

References

-  MESRI, *Plan National pour la Science Ouverte* 2021, (official announcement)
-  EOSC SIRS Task Force, *Scholarly Infrastructures for Research Software* 2020, European Commission, ([10.2777/28598](#))
-  R. Di Cosmo, *Archiving and Referencing Source Code with Software Heritage* ICMS 2020 ([10.1007/978-3-030-52200-1_36](#))
-  R. Di Cosmo, M. Gruenpeter, S. Zacchiroli
Referencing Source Code Artifacts: a Separate Concern in Software Citation,
CiSE 2020 ([10.1109/MCSE.2019.2963148](#)) (hal-02446202)
-  P. Alliez, R. Di Cosmo, B. Guedj, A. Girault, M.-S. Hacid, A. Legrand and N. Rougier
Attributing and referencing (research) software: Best practices and outlook from Inria,
CiSE 2020 ([10.1109/MCSE.2019.2949413](#)) (hal-02135891)
-  J.F. Abramatic, R. Di Cosmo, S. Zacchiroli, *Building the Universal Archive of Source Code*,
CACM, October 2018 ([10.1145/3183558](#))



Appendix

What is at stake: beyond ARDC

Sustainability, technology transfer

Organisational schemas, legal tools, economic models, processes and policies to ensure research software can be maintained and sustained over time, maybe in connection with industry

What is at stake: beyond ARDC

Sustainability, technology transfer

Organisational schemas, legal tools, economic models, processes and policies to ensure research software can be maintained and sustained over time, maybe in connection with industry

Evaluation (funding, careers, etc.)

- avoid the numbers game (beware of *naive software citation counting*)
- identify *roles* in software projects, see:



P. Alliez, R. Di Cosmo, B. Guedj, A. Girault, M.-S. Hacid, A. Legrand and N. Rougier
Attributing and referencing (research) software: Best practices and outlook from Inria,
CiSE 2020 ([10.1109/MCSE.2019.2949413](https://doi.org/10.1109/MCSE.2019.2949413))

What is at stake: beyond ARDC

Sustainability, technology transfer

Organisational schemas, legal tools, economic models, processes and policies to ensure research software can be maintained and sustained over time, maybe in connection with industry

Evaluation (funding, careers, etc.)

- avoid the numbers game (beware of *naive software citation counting*)
- identify *roles* in software projects, see:



P. Alliez, R. Di Cosmo, B. Guedj, A. Girault, M.-S. Hacid, A. Legrand and N. Rougier
Attributing and referencing (research) software: Best practices and outlook from Inria,
CiSE 2020 ([10.1109/MCSE.2019.2949413](https://doi.org/10.1109/MCSE.2019.2949413))

Regulations are coming

software management plans, licensing recommendations, metadata and identification standards

What about FAIR?

FAIR data principles *for data*

in a nutshell: metadata, metadata, metadata all over the place

What about FAIR?

FAIR data principles *for data*

in a nutshell: metadata, metadata, metadata all over the place

But software is *not data* ...

the terms *interoperability* and *reusability* have precise technical meaning for software, and differ significantly from what is intended by the I and R of FAIR;

- see the entries for [software interoperability](#) and [software reusability](#)
- it is *very difficult* to achieve these properties even for commercial software developed by multinationals

What about FAIR?

FAIR data principles *for data*

in a nutshell: metadata, metadata, metadata all over the place

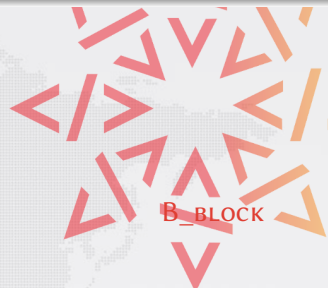
But software is *not data* ...

the terms *interoperability* and *reusability* have precise technical meaning for software, and differ significantly from what is intended by the I and R of FAIR;

- see the entries for [software interoperability](#) and [software reusability](#)
- it is *very difficult* to achieve these properties even for commercial software developed by multinationals

...

let's focus on more actionable issues: Archive, Reference, Describe and Cite are a good starting point



6 SWHIDs by the example

7 News and related initiatives

A word on the trust model for systems of identifiers

Two general classes of systems of identifiers

intrinsic *computed from the object (no registry required, fully decentralised)*
(e.g.: chemical notation, music notation, hashes, SWHIDs)

A word on the trust model for systems of identifiers

Two general classes of systems of identifiers

intrinsic *computed from the object (no registry required, fully decentralised)*
(e.g.: chemical notation, music notation, hashes, SWHIDs)

extrinsic *assigned by an authority (need a registry)*
(e.g.: passport number, DOI, ARK, RRID, etc.)

A word on the trust model for systems of identifiers

Two general classes of systems of identifiers

intrinsic *computed from the object (no registry required, fully decentralised)*
(e.g.: chemical notation, music notation, hashes, SWHIDs)

extrinsic *assigned by an authority (need a registry)*
(e.g.: passport number, DOI, ARK, RRID, etc.)

See [the dedicated blog post](#) for more details

A word on the trust model for systems of identifiers

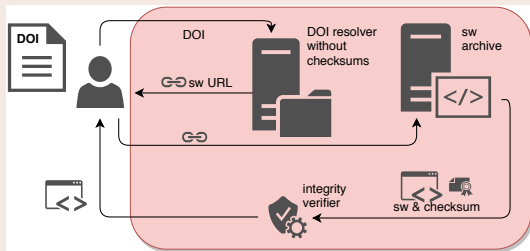
Two general classes of systems of identifiers

intrinsic *computed from the object (no registry required, fully decentralised)*
(e.g.: chemical notation, music notation, hashes, SWHIDs)

extrinsic *assigned by an authority (need a registry)*
(e.g.: passport number, DOI, ARK, RRID, etc.)

See [the dedicated blog post](#) for more details

Trust model, extrinsic (e.g. DOIs)



A word on the trust model for systems of identifiers

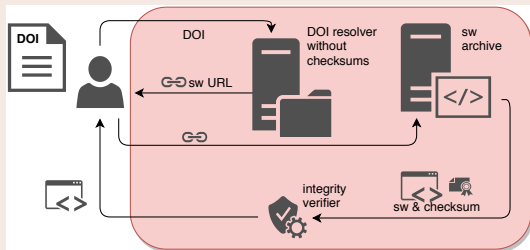
Two general classes of systems of identifiers

intrinsic *computed from the object (no registry required, fully decentralised)*
(e.g.: chemical notation, music notation, hashes, SWHIDs)

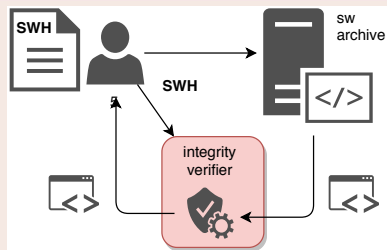
extrinsic *assigned by an authority (need a registry)*
(e.g.: passport number, DOI, ARK, RRID, etc.)

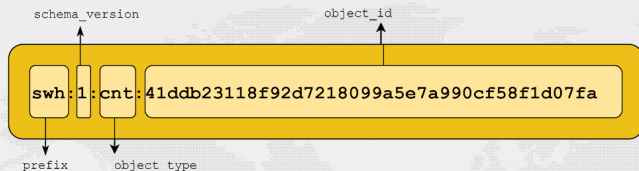
See [the dedicated blog post](#) for more details

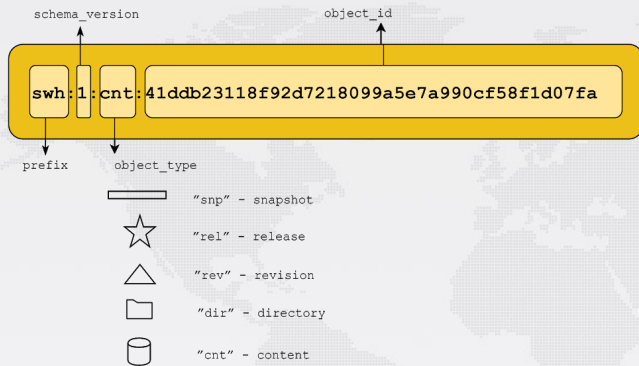
Trust model, extrinsic (e.g. DOIs)



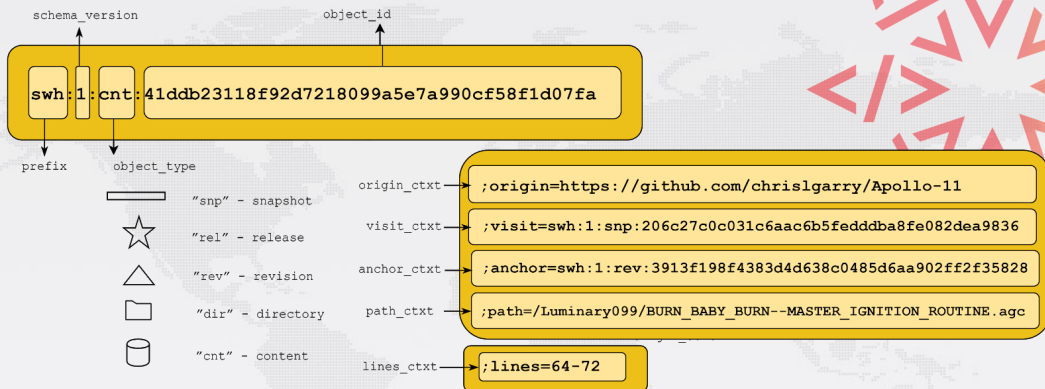
Trust model, intrinsic (e.g. SWHIDs)











An emerging standard

- in Linux Foundation's [SPDX 2.2](#)
- IANA registered, WikiData property [P6138](#)



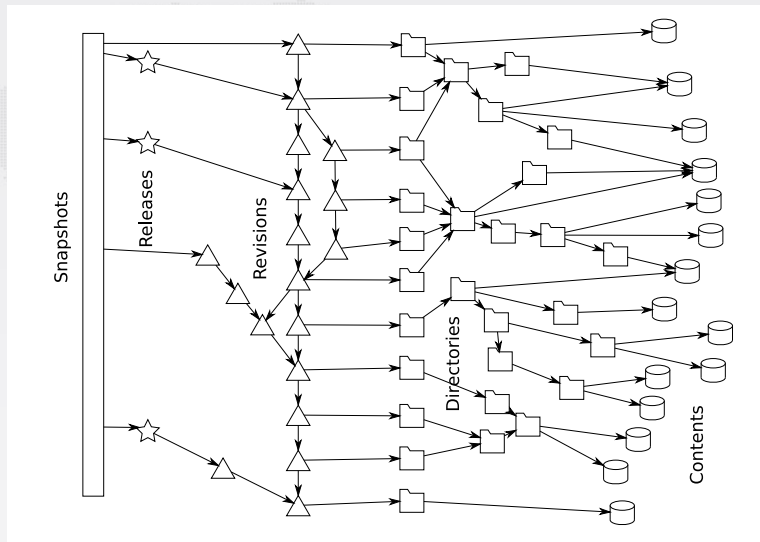
An emerging standard

- in Linux Foundation's [SPDX 2.2](#)
- IANA registered, WikiData property [P6138](#)

Examples:

- [Apollo 11 AGC excerpt](#),
- [Quake III rsqrt](#)

A worked example



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.

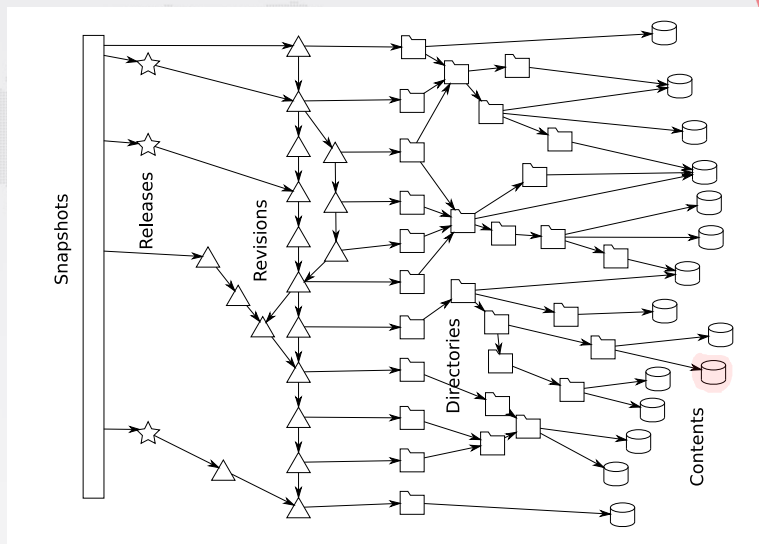
The licenses for most software and other practical works are designed
to take away your freedom to share and change the works. By contrast,
the GNU General Public License is intended to guarantee your freedom to
share and change all versions of a program—to make sure it remains free
software for all 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 things.

To protect your rights, we need to pre

sha1: 8624bcdade55baeef...
sha256: 8ceb4b9ee5aded...
sha1_git: **94a9ed024d385...**
length: 35147

A worked example



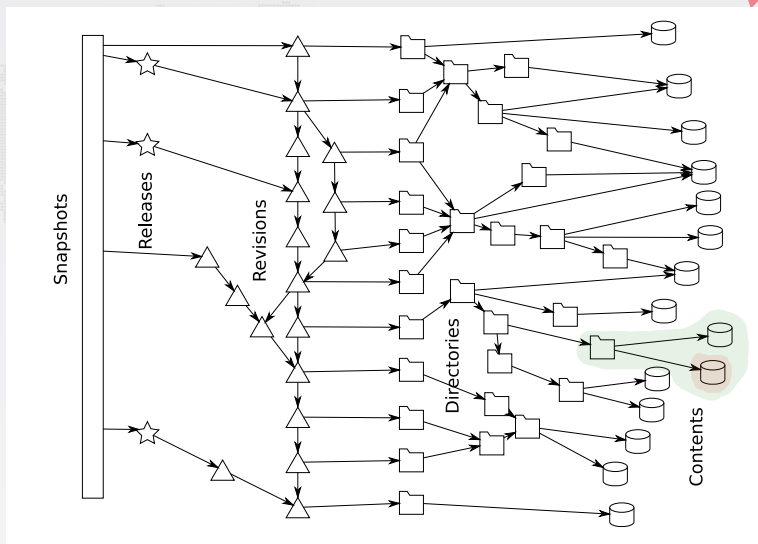


Directories

```
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 e1e10ecef948af0b93adb0372afca89f12e92618a bin
040000 tree 83e56d0beaf7793c77a45a345c80fcb8af503013 debian
040000 tree a34c9c4ba213f0cedc67f9816348d27955577af5 docs
100644 blob f2a6d32c6135aa7287bbd76167b01df2ae4f1539 requirements.txt
100755 blob eee147c36caf1bbcb2d820da8dc026cb5b68180bc setup.py
040000 tree 224bb4c1f4c67fca1d160bffdd2d06094e7e1abf3 sql
040000 tree 8631c9cd77bbe993168107ab5baf51f40c6300be swl
040000 tree 8fb905b56ba8ed692f1209b2773b474c6c1d66c1 utils
```

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d

A worked example



Revisions

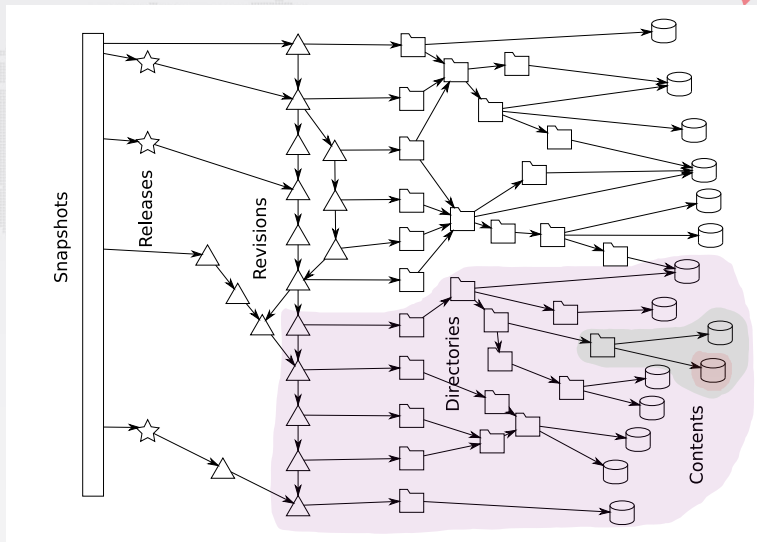
Details	Changes	Files
SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6 Author: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016) Committer: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016) Subject: provenance.tasks: add the revision -> origin cache task Parent: fc3a8b59ca1df424d860f2c29ab07fee4dc35d10 : test..storage: properly pipeline origin and cont... provenance.tasks: add the revision -> origin cache task sw/h/storage/provenance/tasks.py  77		

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

A worked example



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

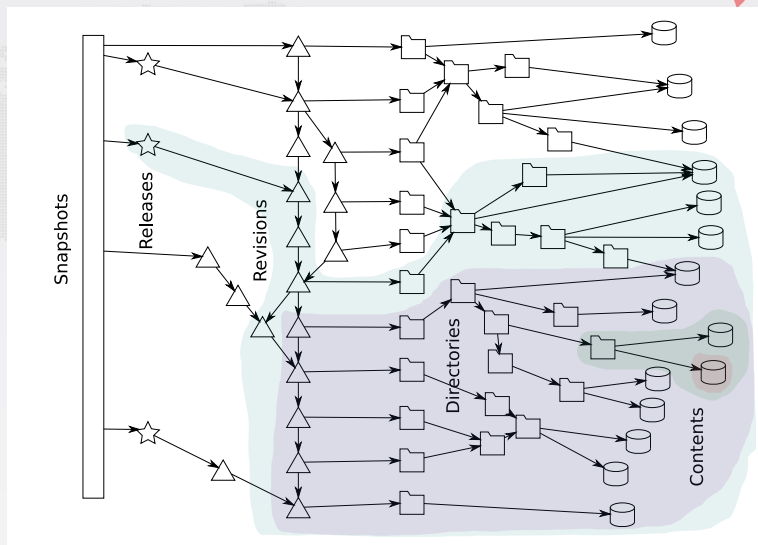
Release swh.storage v0.0.51

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

iQIzBAABCAAdBQJXvZTNFhxuaWNvbGFzQGRhbmRyaW1vbnQuZXUACgkQ7AWLMo2+
neqonw/aq6SOb5DijzEa+kWN3rXgV5+1K1vEVh1wNKAwx8eKJ7aX2kEILDtt7uf
ahpZ6pz3q8nqs6aC1+YrxBfcih3L2YtrdZeWXWqr8xWNMaEoYDb8qaphwh8AD5t2
ICBlit2ujtXuCrDt93eKKPwvzZXg+h80sMWy3SDr6jW7Z7K4Mu/PgGlyLHPY55yo
IGEndWno7VFH1Vm6t1n5qB7I5mXRaqA+becqddubTZ2xij+jpIUqC8cyqN3hm/fL
qsJ2mu8kyz3t8tG/H1/pV+I5OwBlNpO5STH0tUjoEVgPK/dH5P79QuHDH2FkCao
klj6kAWyU80Mxb+nKV/jelBrR3+yWBFJ3Qp5a1/V8oOTb6E1dALcNMpEaKCoKtMt
d/gMRax11I1/g0EDfnsW67G6sDwKPKPHngfVLQ3nV3GaQQTnu1RpMz006H9/tAwzC
Gg/K1PdHT4hz0I46wYPZyje0U2VXGFu6vVU9vFQ4ZR/Wjn+0zmZdcRdrJJSUOMN
RpTTFusbXUeXHGOpkgXhSYTnvp1gdPc76U5TsK0aGe84AZm1Ik0mGrwXCVfPqIYo
nhhibB5HBNMoqyF6yTSOpUbYK70tpYRRUGKwDeRK0wKSxkWUzGtKzy6jYqJjo29
gulwgZQif5qWQCB00ontAL2+HvPfFaVyckMejUhg62cP/+EHlvUk=
=kOxP
---END PGP SIGNATURE---

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb

A worked example

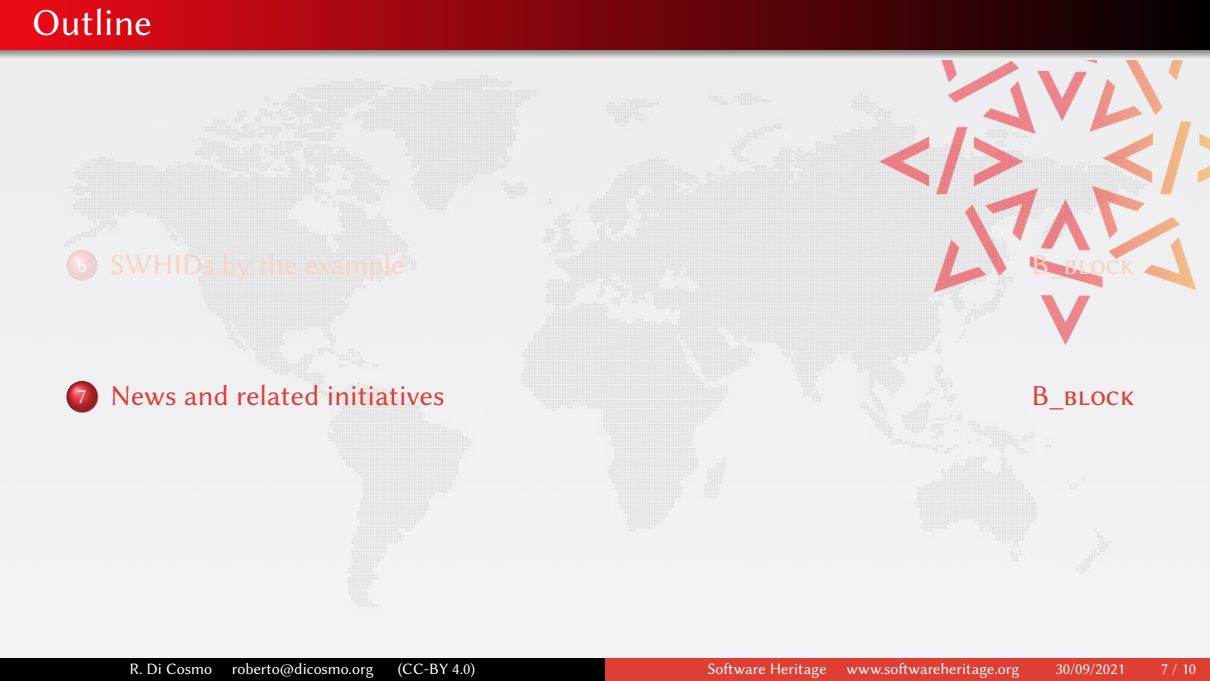


git show-refs

Snapshots

```
commit 08ffeb25770109525eb3ce21691466c53a1d9158 refs/heads/atime
commit ba5443a24e3f9fe323a46c292cec4fcbe61c67eb refs/heads/directory-listing-arrays
commit d69e0dbf892383ff6589b27fbc1c05d27238d9c5 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 586fba4e580b4f5fab05f599367643cbb1a9c7f refs/tags/v0.0.15
tag 8cd8b885f4080bf363177742bd289f660e5be51c refs/tags/v0.0.16
tag a542444ee3f0fbcd35efb202fee035c809abc7d6 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



6 SWHIDs by the example

7 News and related initiatives

B_BLOCK

B_BLOCK

Recent news, and a lesson to be learned

Saving 250.000 endangered repositories...

- summer 2019: BitBucket announce Mercurial VCS phase out
- fall 2019: Software Heritage teams up with Octobus (funded by NLNet, thanks!)
- july 2020: BitBucket erases *250.000* repositories
- august 2020: bitbucket-archive.softwareheritage.org is live

Recent news, and a lesson to be learned

Saving 250.000 endangered repositories...

- summer 2019: BitBucket announce Mercurial VCS phase out
- fall 2019: Software Heritage teams up with Octobus (funded by NLNet, thanks!)
- july 2020: BitBucket erases 250.000 repositories
- august 2020: bitbucket-archive.softwareheritage.org is live

... preserving the web of knowledge

([Tweet is here](#))



Gabriel Altay
@gabrielaltay

Just realized @Bitbucket disabled all mercurial repositories when the @asclnet informed me that a link associated with an old paper of mine was down. Thought all was lost, but someone archived all the repos! very classy move by @octobus_net and @SWHeritage.

[Traduire le Tweet](#)

1:48 AM · 31 août 2020 · Twitter Web App

Bottomline

explicit deposit is important, ...

... and we must promote it...

... but will never be enough.

(think also of all software dependencies!)

Thomas Jefferson, February 18, 1791

...let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.

Thomas Jefferson, February 18, 1791

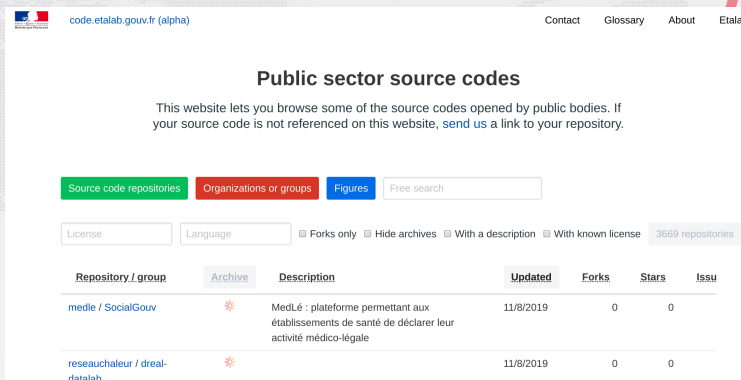
...let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.

Welcoming ENEA



Italian National Agency for New Technologies,
Energy and Sustainable Economic Development

- first **institutional** mirror
- increased resilience
- **AI infrastructure** for researchers
- stepping stone to
an European joint effort



code.etalab.gouv.fr (alpha)

Contact Glossary About Etalab

Public sector source codes

This website lets you browse some of the source codes opened by public bodies. If your source code is not referenced on this website, [send us](#) a link to your repository.

Source code repositories Organizations or groups Figures Free search

License Language Forks only Hide archives With a description With known license 3669 repositories

Repository / group	Archive	Description	Updated	Forks	Stars	Issu
medle / SocialGouv	✱	MedLé : plateforme permettant aux établissements de santé de déclarer leur activité médico-légale	11/8/2019	0	0	
reseauchaleur / dreai-datalab	✱		11/8/2019	0	0	

code.etalab.gouv.fr (alpha)

Contact Glossary About Etalab

Public sector source codes

This website lets you browse some of the source codes opened by public bodies. If your source code is not referenced on this website, [send us](#) a link to your repository.

Source code repositories Organizations or groups Figures Free search

License Language ☐ Forks only ☐ Hide archives ☐ With a description ☐ With known license 3669 repositories

Repository / group	Archive	Description	Updated	Forks	Stars	Issu
medle / SocialGouv	✳	MedLé : plateforme permettant aux établissements de santé de déclarer leur activité médico-légale	11/8/2019	0	0	
reseauchaleur / dreal-datalab	✳		11/8/2019	0	0	

<https://code.etalab.gouv.fr>

Paris Call on Software Source Code

“[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”

Paris Call on Software Source Code

“[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”

SWHAP : an important step forward

- detailed guidelines to **curate** landmark legacy source code and **archive** it on Software Heritage
- intense cooperation with **Università di Pisa** and **UNESCO**
- open to all, we'll promote it worldwide

<https://www.softwareheritage.org/swhap>