

Software Heritage: building the software pillar of Open Science

a stepping stone for reproducibility

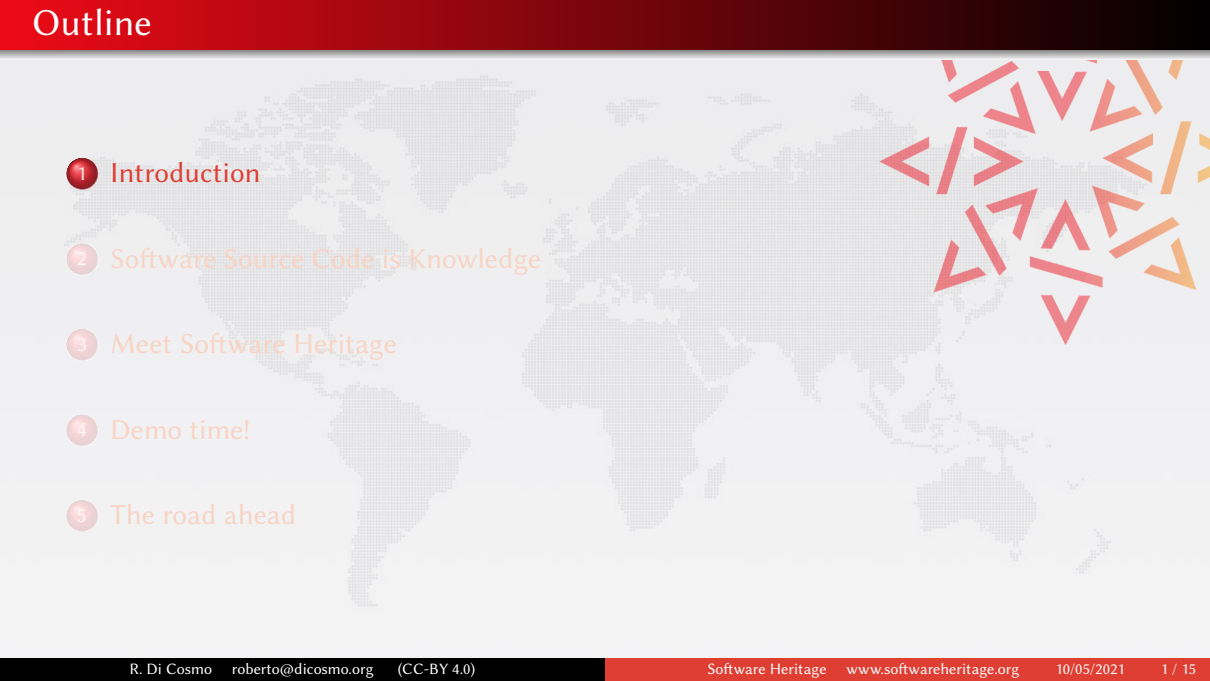
Roberto Di Cosmo
Inria and Université de Paris

Journée reproductibilité



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.”

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

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

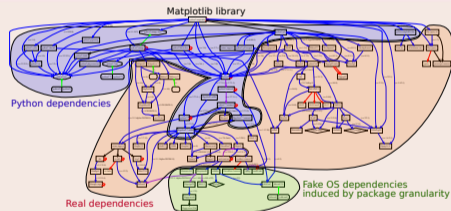
Source code is *special* (software is *not* data)

Software *evolves* over time

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

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*

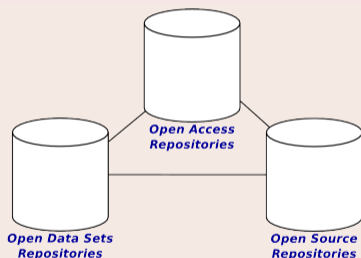


Precious, endangered *executable* and *human readable* knowledge

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

no organised effort to catalog and archive it

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*)

Reference

Research software artifacts must be properly **referenced**
make sure we can *identify* them (*reproducibility*)

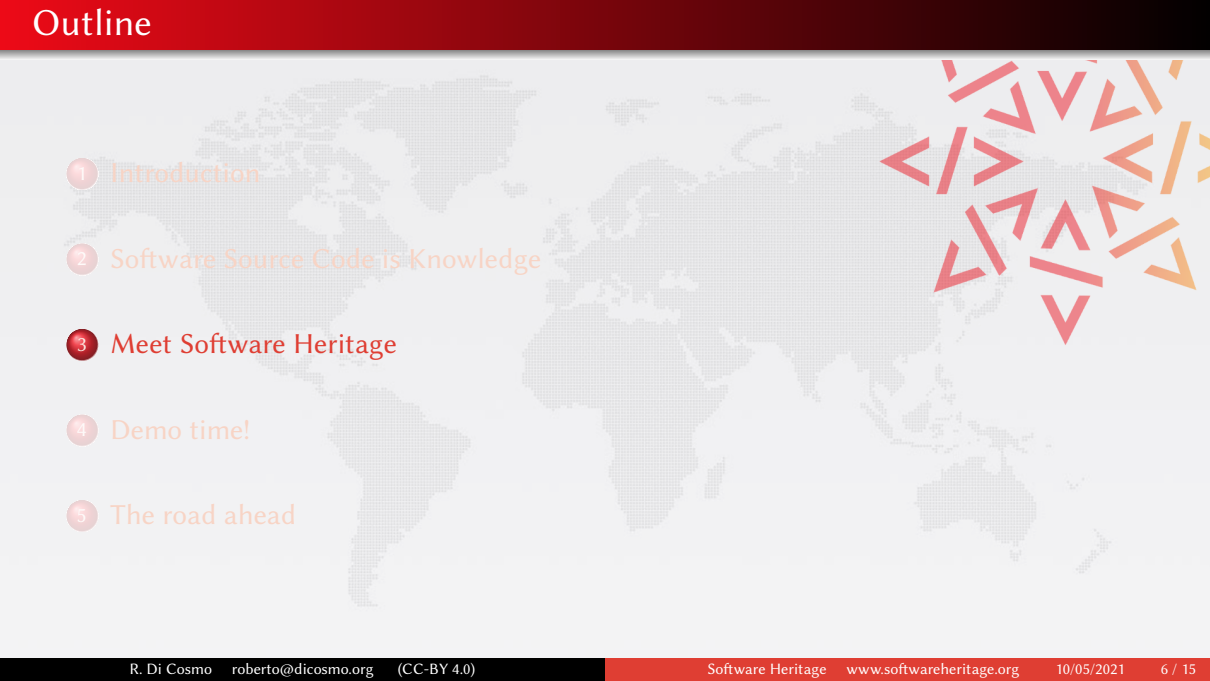
Describe

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

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

Cultural Heritage



Industry



Research



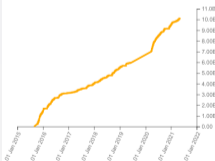
Education



Software Heritage

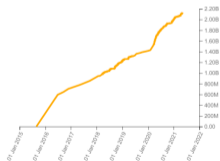
Source files

10,162,254,604



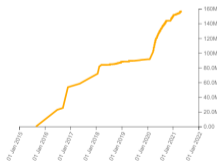
Commits

2,128,899,595



Projects

156,417,833



Directories

8,477,359,190

Authors

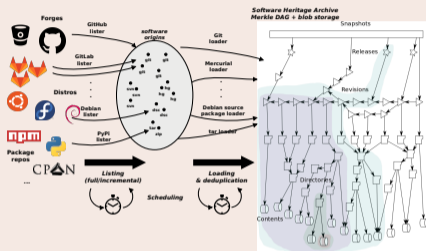
42,598,760

Releases

17,293,952

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)

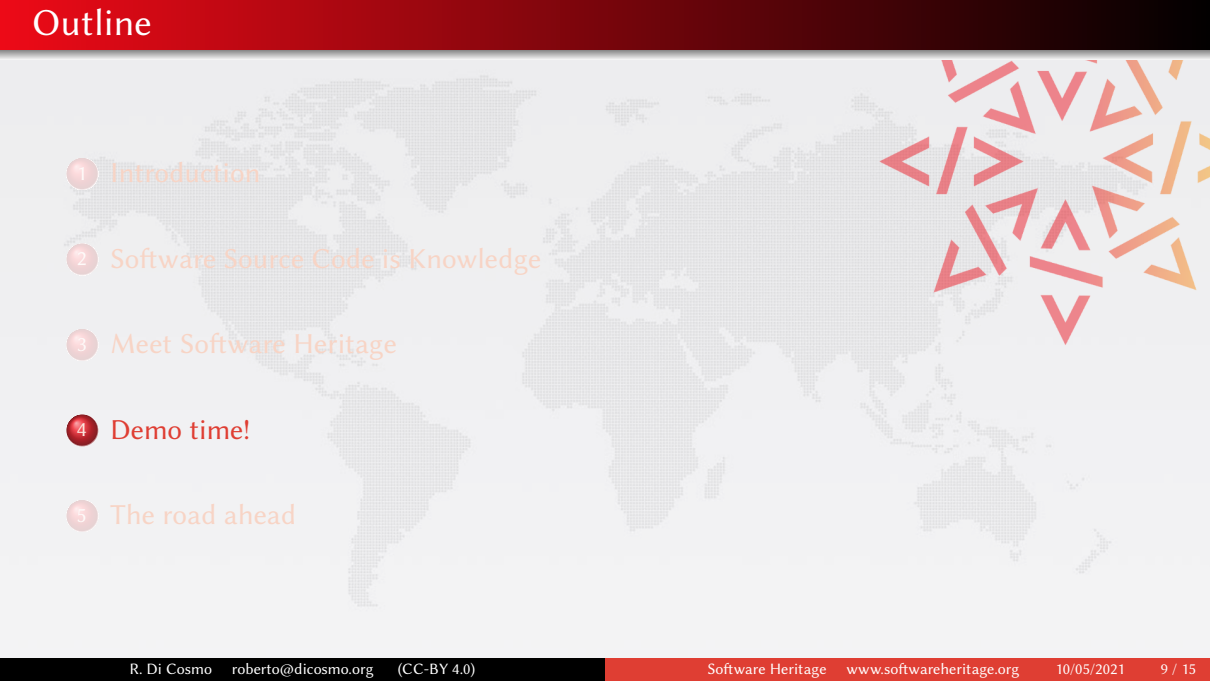
Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs



Now supported in SPDX 2.2, Wikidata etc.

Cite/Credit

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

- 
- 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 use in a research article: compare Fig. 1 and conclusions
 - in [the 2012 version](#)
 - in [the updated version](#) using SWHIDs and Software Heritage
- Cite software [with the biblatex-software style](#) from CTAN
- Example use in a research article: extensive use of SWHIDs in [a replication experiment](#)
- Example in a real journal: [an article from IPOL](#)
- Supporting reproducible builds: [Guix](#) and [Nix](#)

Saving 250.000 endangered repositories...

- summer 2019: BitBucket announce Mercurial VCS phase out
- fall 2019: Software Heritage teams up with Octopus (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 [@octopus_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!)

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

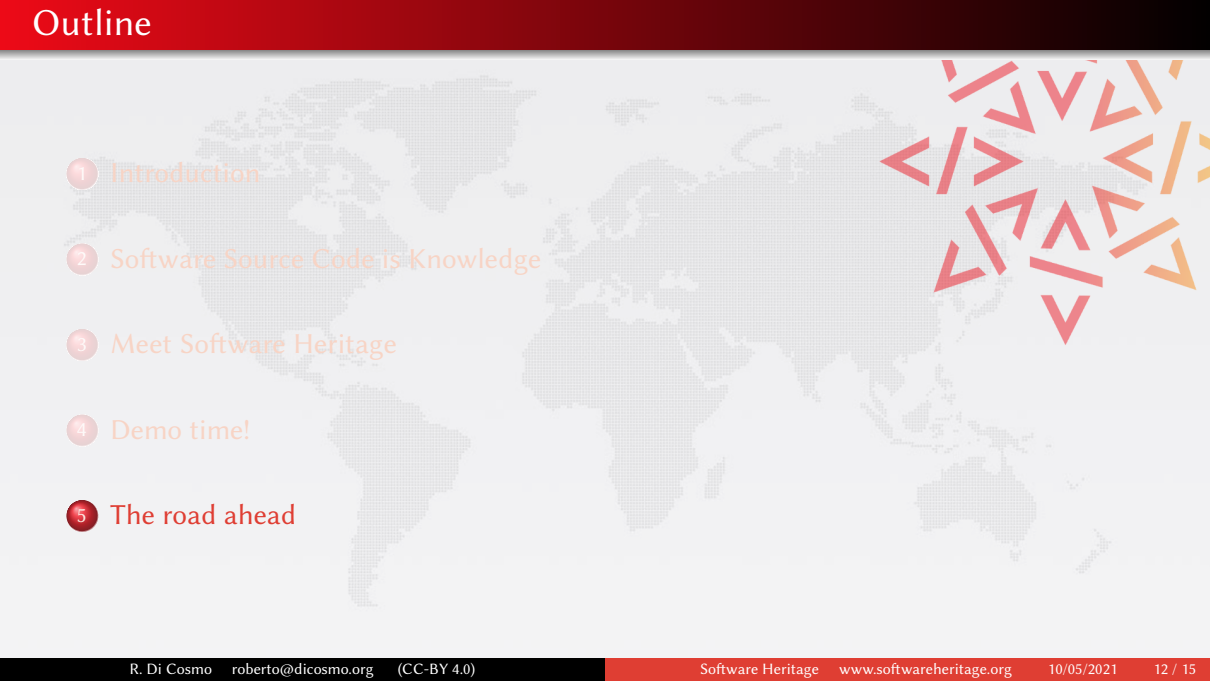
SWHID standard is independent of version control systems

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

Donors, members, sponsors



INVENTEURS DU MONDE NUMÉRIQUE

Platinum sponsors



MINISTÈRE
DE L'ENSEIGNEMENT SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION



HUAWEI



SOCIÉTÉ
GÉNÉRALE

Gold sponsors



Université
de Paris



SORBONNE
UNIVERSITÉ

Silver sponsors



Software Intelligence for Digital Leaders



Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE



UNIVERSITÀ
DI PISA



Bronze sponsors



DANS



FOSSID



Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

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 (academic, software & tutorial files)
- 2 Save your code (<http://osdn.sourceforge.net/>)
- 3 Reference your work (full repository, specific version or code fragments)

- summary
- ICMS 2020

You may help!

Foster adoption and best practices

- archive and reference your source code, cite using [biblatex-software](#)
- describe and deposit via [HAL](#) (see Morane Gruenpeter's talk later today)
- ask journals and conferences to use Software Heritage (like [IPOL](#), etc.)

Engage with Software Heritage as an individual

- the [ambassador program](#) is open:
 - learn directly from the core team, share with your research community
- (small) [grants are available](#) to expand the coverage of the archive

Engage with Software Heritage as an organization

- join as a [member/sponsor](#)
- establish a [mirror](#) (see e.g. [ENEA](#)): support the mission and enable Big Code

Questions?

References

-  R. Di Cosmo, *A revolutionary infrastructure for Open Source*, 2021, EU Software Forum (slides) (video)
-  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. Zacchioli
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. Zacchioli, *Building the Universal Archive of Source Code*,
CACM, October 2018 (10.1145/3183558)