

# Towards a Software Pillar for Open Science policy and implementation

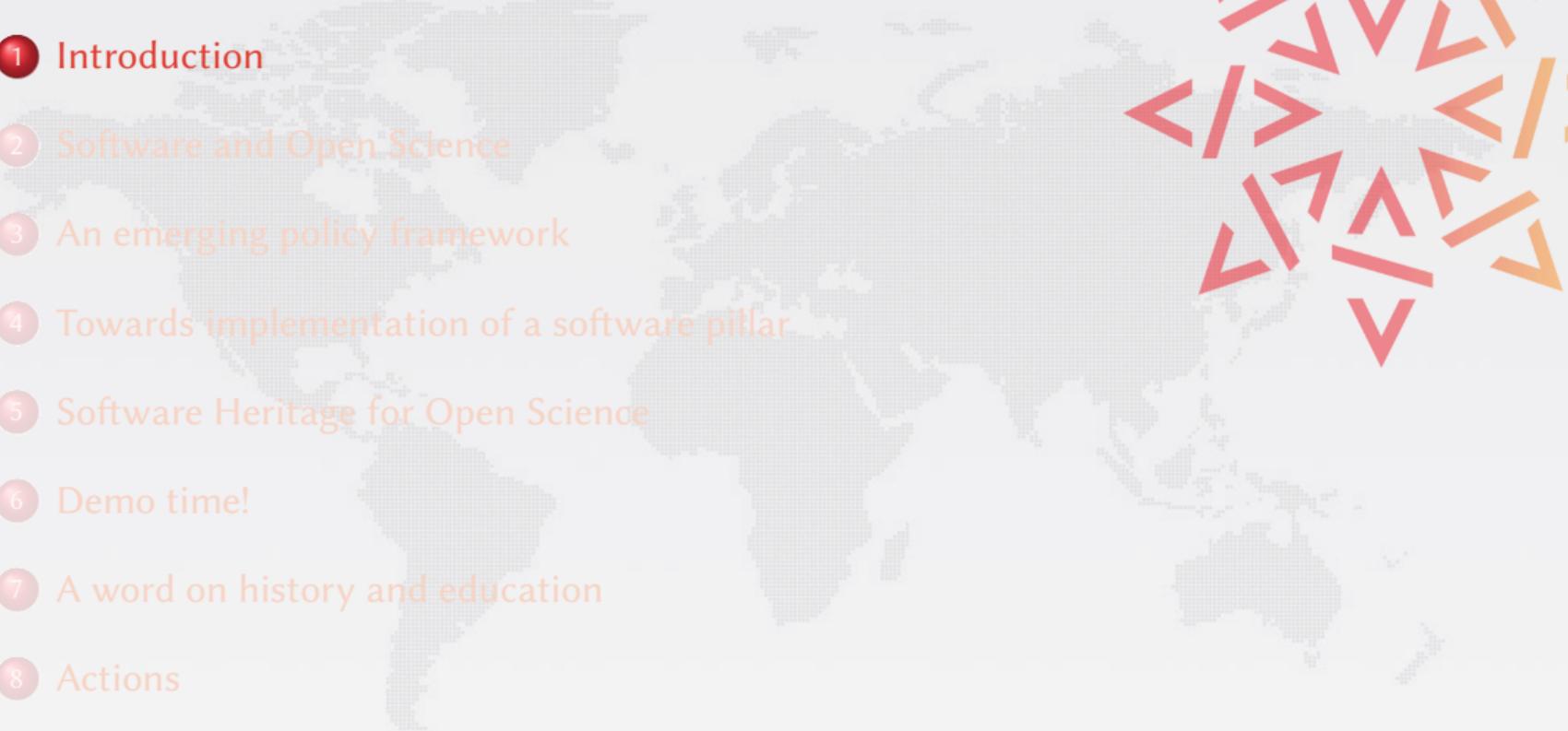
Roberto Di Cosmo  
Biennale Tecnologia, Torino

Director, Software Heritage  
Inria and Université de Paris Cité

November 12th 2022



Software Heritage  
THE GREAT LIBRARY OF SOURCE CODE

- 
- 1 Introduction
  - 2 Software and Open Science
  - 3 An emerging policy framework
  - 4 Towards implementation of a software pillar
  - 5 Software Heritage for Open Science
  - 6 Demo time!
  - 7 A word on history and education
  - 8 Actions



# Short Bio: Roberto Di Cosmo

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](http://www.mancoosi.org)

2010 *IRILL* [www.irill.org](http://www.irill.org)

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science, France*

2021 *EOSC Task Force on Infrastructures for Software,  
European Union*

- 
- 
- 1 Introduction
  - 2 Software and Open Science
  - 3 An emerging policy framework
  - 4 Towards implementation of a software pillar
  - 5 Software Heritage for Open Science
  - 6 Demo time!
  - 7 A word on history and education
  - 8 Actions

# Why Open Science?

Open Science ([Second National Plan for Open Science](#), France, 2021)

*Unhindered dissemination of results, methods and products from scientific research. It draws on the opportunity provided by recent digital progress to develop open access to publications and – as much as possible – data, source code and research methods.*

Jean-Eric Paquet (EU DGRI, [on the objective of Open Science](#))

*“Increase scientific quality, the pace of discovery and technological development, as well as societal trust in science.”*

Mariya Gabriel ([EU Commissioner for Research](#))

The COVID-19 crisis has also shown that cooperation at international level in research and innovation is more important than ever, including through *open access to data and results. No nation, no country can tackle any of these global challenges alone.*

Yuval Noah Harari (on COVID 19)

*“The real antidote [to epidemic] is scientific knowledge and global cooperation.”*

# Two well known pillars of Open Science

## Open Access (long, painful, unfinished story)

19XX's compulsory exclusive copyright transfer to publishers

(notable exceptions: [US federal agencies](#) and [UK Crown Copyright](#))

1990's Internet, Web and ArXiv break the [marriage of convenience of researchers with publishers](#)

2010's reactions (SciHub, 2011; [Plan S](#), 2018) and transformations ([not so easy](#))

TL;DR: see [my viewpoint in 2005](#) and the [SIGPLAN blog in 2020](#)

## Open Data (less painful, but still unfinished story)

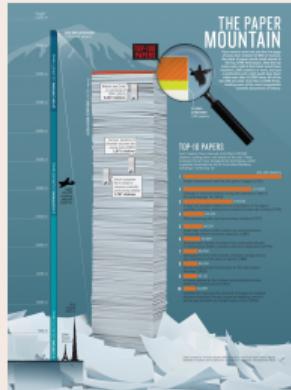
- 2006 (and 2021): OECD recommendation on [publicly funded research data](#)
- 2016 and later: FAIR terminology (*focus on metadata, sort of forgets open...*)

## Breaking news, August 25, 2022: [US OSTP memo to federal agencies](#)

**zero embargo** on public access to publications and data

# Software: the third pillar of Open Science

Software powers modern research



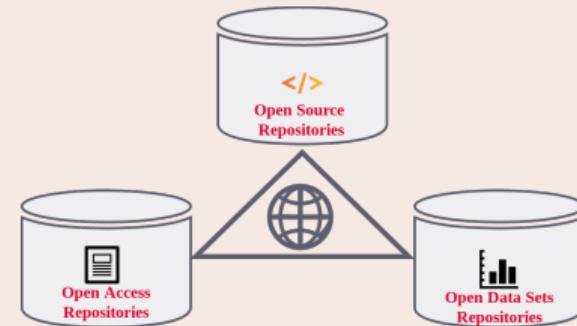
[...] software [...] essential in their fields.

*Top 100 papers (Nature, 2014)*

*Sometimes, if you dont have the software, you dont have the data*

*Christine Borgman, Paris, 2018*

A key pillar: software (source code)



The links in the picture are **important**

Nota Bene

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

access to the *source code* is essential!

Preserving (the history of) source code is necessary for *reproducibility*

# 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     G0PERF1
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;

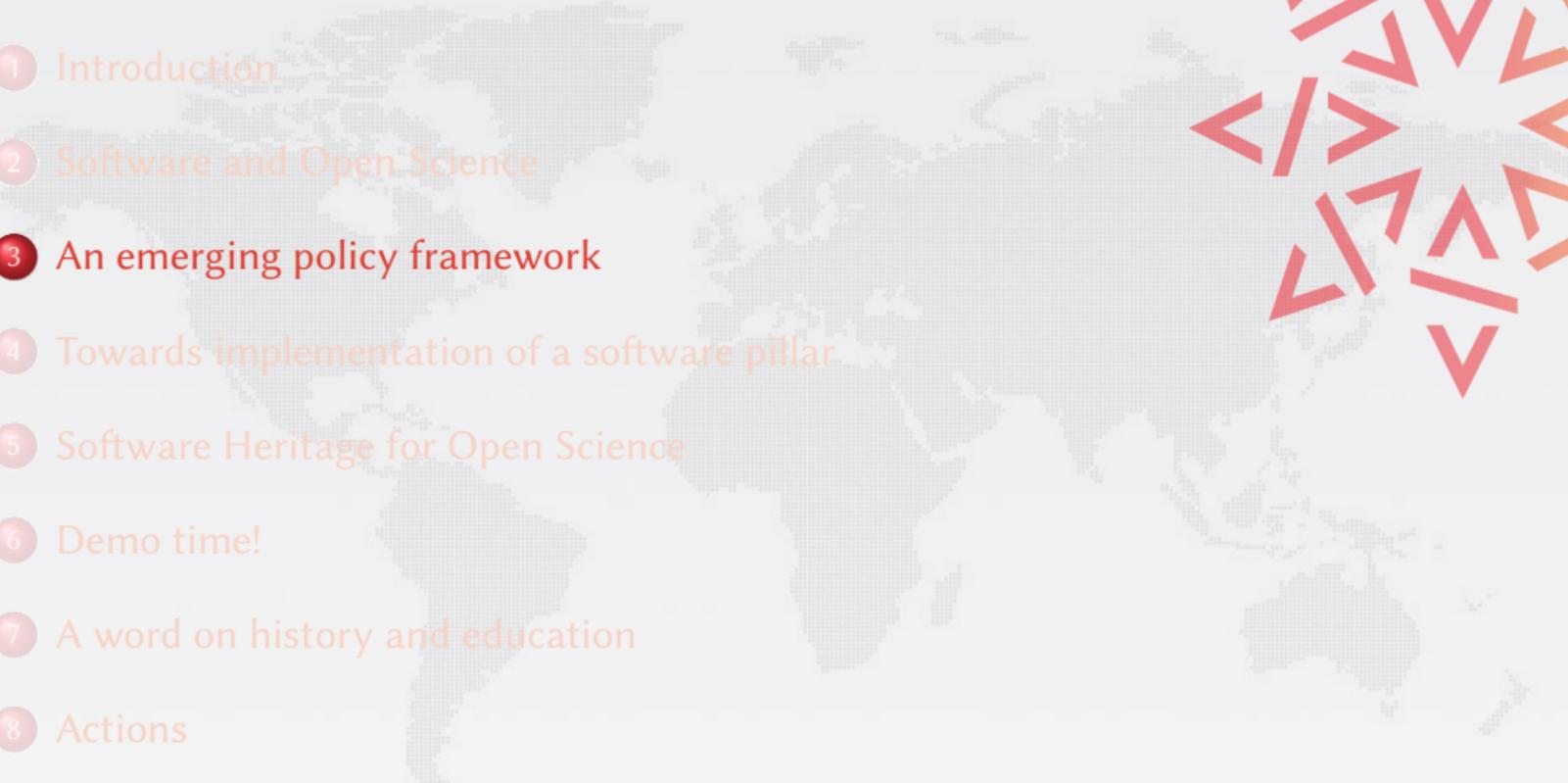
    x2 = number * 0.5F;
    y = number;
    i = *( long * ) &y; // evil floating point bit level hacking
    i = 0x5f3759df - ( i >> 1 ); // what the fuck?
    y = * ( float * ) &i;
    y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
// y = y * ( threehalfs - ( x2 * y * y ) ); // 2nd iteration,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

    return y;
}
```

Len Shustek, Computer History Museum

2006

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

- 
- 
- 1 Introduction
  - 2 Software and Open Science
  - 3 An emerging policy framework
  - 4 Towards implementation of a software pillar
  - 5 Software Heritage for Open Science
  - 6 Demo time!
  - 7 A word on history and education
  - 8 Actions

# International highlights

## Paris Call on Software Source code (2019, UNESCO)



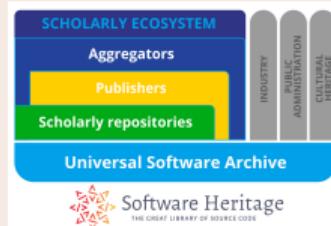
40 international experts call to “promote software development as a valuable research activity, and research software as a key enabler for Open Science/Open Research, [...] recognising in the careers of academics their contributions to high quality software development, in all their forms”

## UNESCO recommendations for Open Science, 2018-2021

*“The source code must be included in the software release and [...] the license must allow modifications, derivative works and sharing [...]”*

*“Open science infrastructures should be [...] essentially not-for-profit and long-term”*

## EOSC SIRS report: Software Source Code and Open Science, 2020



- connect scholarly ecosystem via Software Heritage
- use open non profit infrastructures
- open source first: *“all research software should be made available under an Open Source license by default”*

# French National plan for Open Science, 2021-2024

MINISTÈRE  
DE L'ENSEIGNEMENT  
SUPÉRIEUR,  
DE LA RECHERCHE  
ET DE L'INNOVATION  
L'enseignement supérieur et la recherche  
Ministère de l'Education nationale et de la recherche

## Second French Plan for Open Science



### Path Three : Opening up and promoting source code produced by research

7

Recognize and support the dissemination under an open source license of software produced by publicly funded research programmes

8

Highlight the production of source code from higher education, research and innovation

9

Define and promote an open source software policy

« The opening of software source code is a major challenge for the **reproducibility** of scientific results. »

« Distribution of software products under **open source licence** will be preferred. »



Accueil > Recherche > Science ouverte

Publié le 05.02.2022

#### Sommaire

- The Coq proof assistant : lauréat de la catégorie Scientifique et technique
- SciLifeLab : lauréat de la catégorie Recherche fondamentale
- Faust : lauréat de la catégorie Documentation
- GammaRay : prix du jury
- Jury

### Remise des prix science ouverte du logiciel libre de la recherche

Le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation remet pour la première année les Prix science ouverte du logiciel libre de la recherche. Dix logiciels mis au point par des équipes françaises sont récompensés pour leur contribution à l'avancée de la connaissance scientifique.

- 129 projects
- 4 awards
- 6 accessit
- first edition

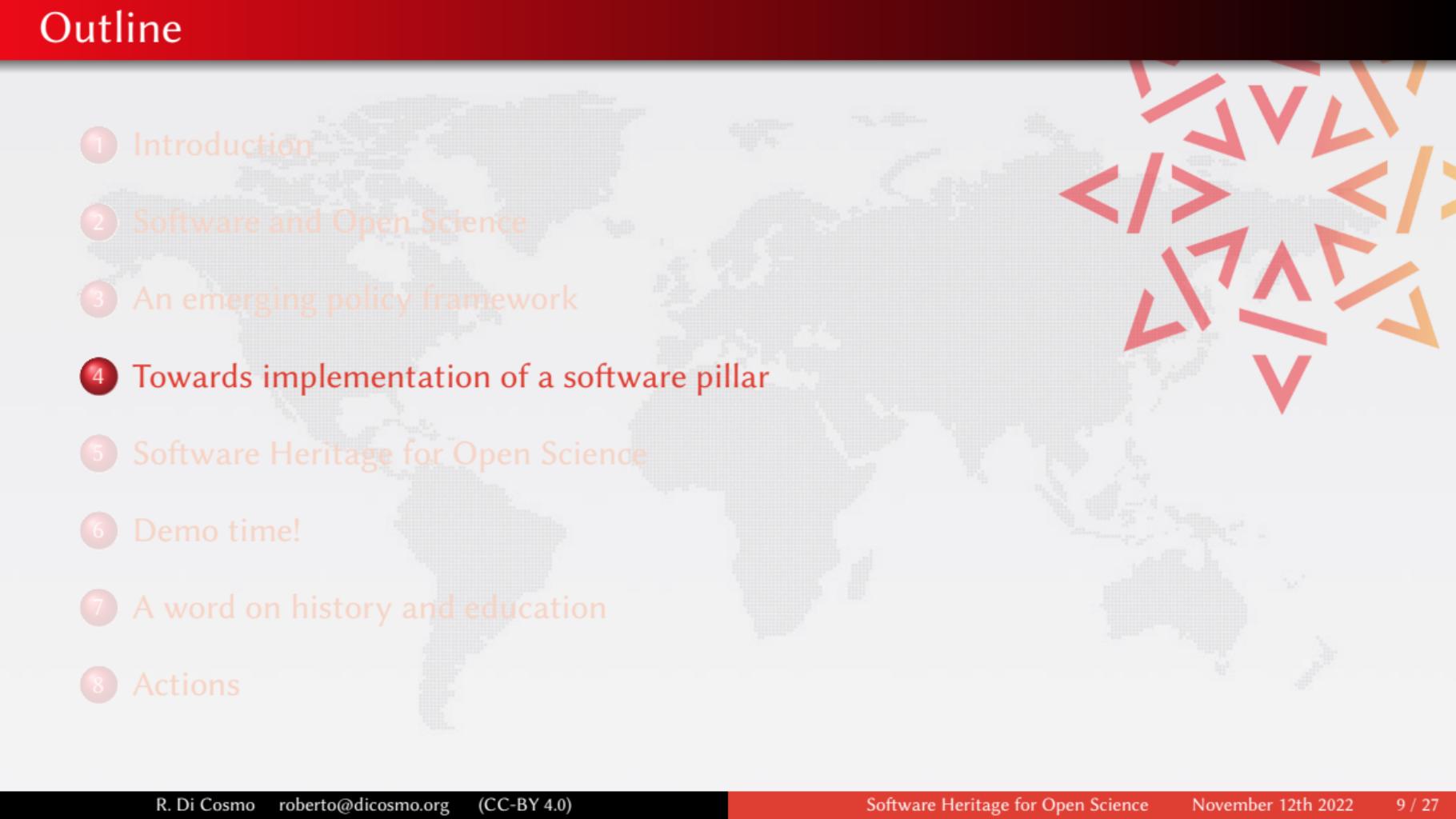
## Five action lines

- Identifying and highlighting research software production
- Technical and social tools and best practices
- Valorization and sustainability
- Liaison and animation at national, European, and international levels
- Recognition and careers

## Leveraging experience and connections

- Open Source thematic group in Systematic (since 2007, more on demand)
- Collaboration with DINUM, Eclipse Foundation, OW2, ...

# Outline

- 
- 1 Introduction
  - 2 Software and Open Science
  - 3 An emerging policy framework
  - 4 Towards implementation of a software pillar
  - 5 Software Heritage for Open Science
  - 6 Demo time!
  - 7 A word on history and education
  - 8 Actions

# A plurality of needs

## Researchers

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

## Laboratories/teams

- track software contributions
- produce reports
- maintain web page

## Research Organization

know its software assets

- technology transfer
- impact metrics
- funding strategy
- career evaluation

# What is at stake

## ARDC

- **Archive** for retrieval (*reproducibility*)
- **Reference** for identification (*reproducibility*)
- **Describe** for discovery and reuse
- **Cite/Credit** for credit and evaluation

## Before ARDC

- **Development** practices and tools (VCS, build system, test suites, CI, ...)
- **Opening up** towards a community (documentation, organization, communication)

Need training, best practices

## Beyond ARDC

- **Policies** (dissemination, reuse, careers!)
- **Sustainability** (legal, economic etc.)
- Technology transfer
- Advanced technologies and tools (quality, traceability, etc.)

let's focus on infrastructures for ARDC

# Forges are *not* archives!

## 2015: the first big bad news

Google Code and Gitorious.org shutdown: ~1M endangered repositories

- broken links in the web of knowledge (my papers too)

## Big bad news keep coming in

- summer 2019: BitBucket announces Mercurial VCS sunset
- july 2020: BitBucket erases 250.000+ repositories (including research software)
- summer 2022: GitLab.com considers erasing **all** projects that are **inactive for a year**

## In Academia too!

- 2021: Inria's old gforge is unplugged... **breaks the Opam build chain** for OCaml

We need a universal archive of software source code: now we have one!

- 
- 1 Introduction
  - 2 Software and Open Science
  - 3 An emerging policy framework
  - 4 Towards implementation of a software pillar
  - 5 Software Heritage for Open Science
  - 6 Demo time!
  - 7 A word on history and education
  - 8 Actions



# 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

damage  
disaster  
media  
attack  
aging  
obsoletedependencies  
malicious  
dangling  
reference  
deletion  
corruption  
storage  
encryption  
format

preserve and share all software source code

## Research infrastructure



enable analysis of all software source code

# The largest software archive, a shared infrastructure



Bitbucket	1,925,997 origins	git	21,603 origins	R	21,113 origins
debian	128,719 origins	Guix	5,947 origins	GitHub	137,564,899 origins
GitLab	3,982,586 origins	GNU	12,032 origins	Maven	354 origins
heptapod	1,068 origins	launchpad	329,908 origins	NixOS	93,738 origins
Phabricator	192 origins	npm	1,799,296 origins	npm	4,083 origins
		python	410,582 origins	SOURCEFORGE	308,990 origins

## Sharing the vision



United Nations  
Educational, Scientific and  
Cultural Organization



And many more ...

[www.softwareheritage.org/support/testimonials](http://www.softwareheritage.org/support/testimonials)

## Donors, members, sponsors



Diamond sponsor



Platinum sponsors



Gold sponsors



Silver sponsors

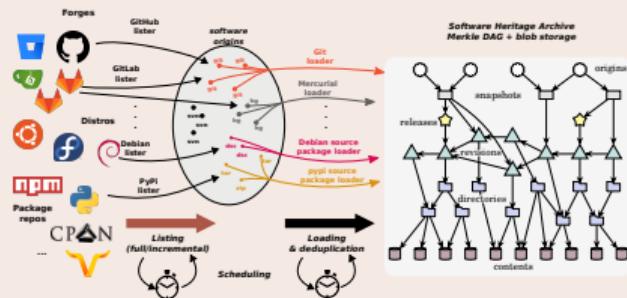


Bronze sponsors



# Addressing the four ARDC needs (see ICMS 2020 for details)

## Archive (12B+ files, 180M+ projects)



- [save.softwareheritage.org](https://save.softwareheritage.org)
- [deposit.softwareheritage.org](https://deposit.softwareheritage.org)

## Reference (25 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs



Now supported in SPDX 2.2, Wikidata etc.

## Describe

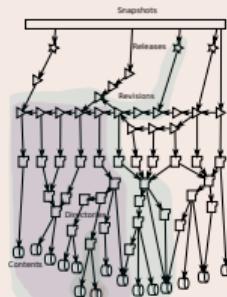
- *Intrinsic metadata* from source code
- Contributed the [Codemeta generator](#)

## Cite/Credit

- Contributed [software citation style](#) [biblatex-software](#), v 1.2-2 now on CTAN

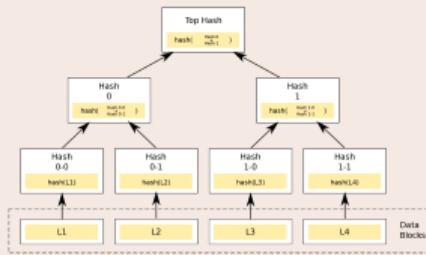
# A revolutionary infrastructure

## The *graph* of Software Development



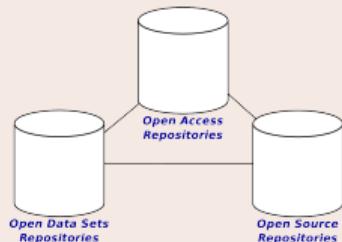
All software development  
in a single graph ...

## The *blockchain* of Software Development



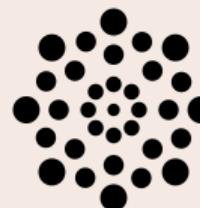
... a single  
Merkle graph!

## A pillar of Open Science



Reference **archive** of  
Research Software

## Reference platform for *Big Code*



A **single, uniform** data struc-  
ture

Thomas Jefferson, February 18, 1791

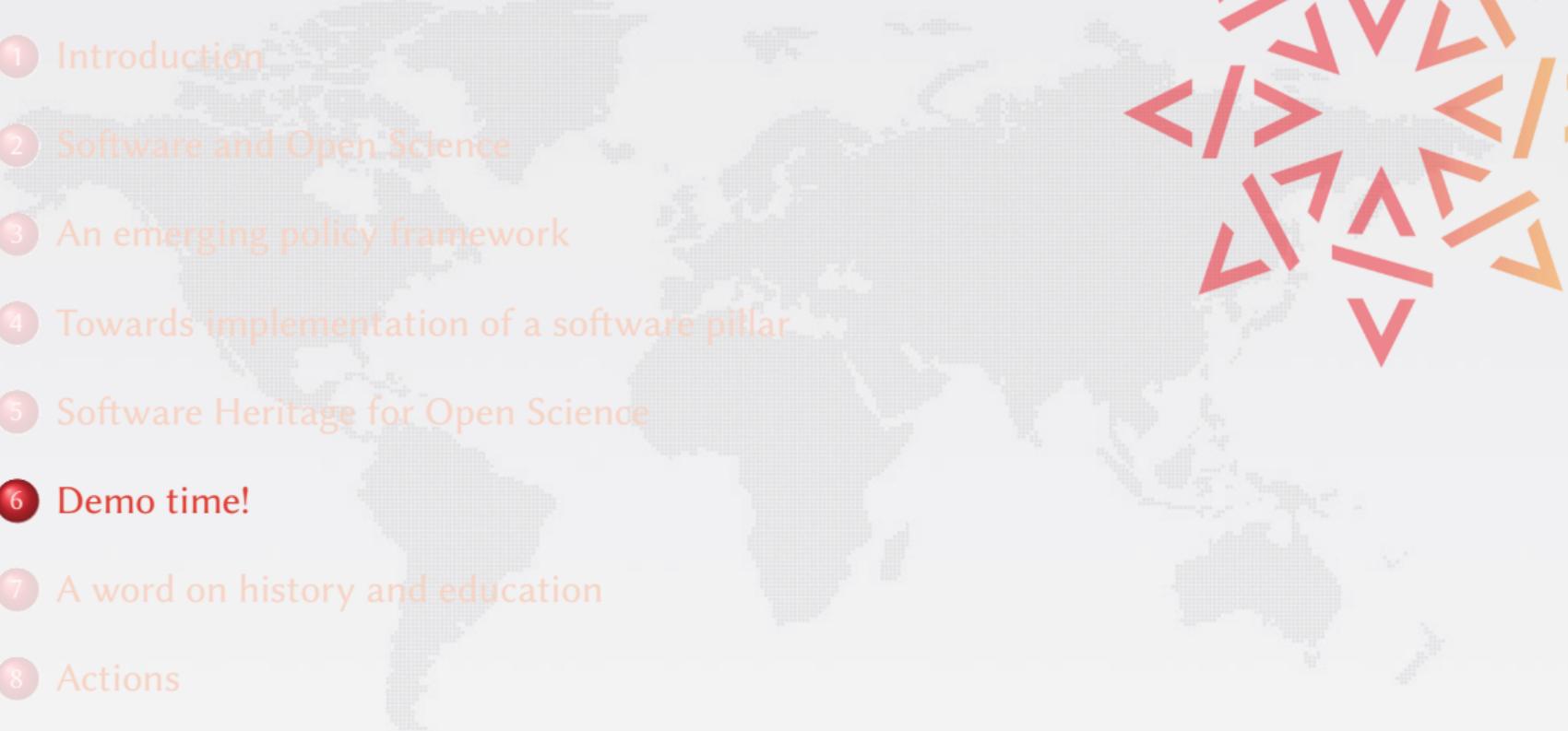
*...let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.*

First institutional mirror will be in Italy



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

- agreement in 2019
- deployment ongoing
- stepping stone to  
an European joint effort

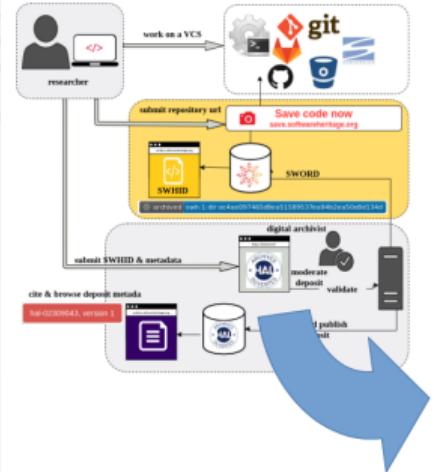
- 
- 1 Introduction
  - 2 Software and Open Science
  - 3 An emerging policy framework
  - 4 Towards implementation of a software pillar
  - 5 Software Heritage for Open Science
  - 6 Demo time!
  - 7 A word on history and education
  - 8 Actions

# A walkthrough

- Browse (e.g. [Apollo 11](#), and your work may be already there !)
- Trigger archival, use the [updateswh](#) browser extension (GitHub action available too)
- Get and use SWHIDs ([full specification available online](#))
- Cite software with [biblatex-software](#) package from CTAN
  - [Overleaf ACMART template](#) available
- Example in journals: [article from IPOL](#)
- Example with Parmap: [devel on Github](#), [archive in SWH](#), [curated deposit in HAL](#)
- Extracting all the software products [for Inria](#), [for CNRS](#), [for CNES](#), [for LIRMM](#) or [for Rémi Gribonval](#) using [HalTools](#)
- Curated deposit in SWH via HAL, see for example: [LinBox](#), [SLALOM](#), [Givaro](#), [NS2DDV](#), [SumGra](#), [Coq proof](#), ...
- Example use in research articles:
  - compare Fig. 1 and conclusions in [the 2012 version](#) and [the updated version](#)
  - SWHID in [a replication experiment](#)



# Overview of the Software Heritage / HAL synergy



<https://hal.archives-ouvertes.fr/hal-02130801>

This screenshot shows the HAL archive page for the identifier [hal-02130801](https://hal.archives-ouvertes.fr/hal-02130801). The page includes a logo for HAL open science, a search bar, and navigation links for Home, Submit, Browse, Search, and Documentation. The main content area displays the title "LinBox", version 1.3.3, and a brief description of the library. It lists several institutions involved in its development, including INRIA Grenoble - Rhône-Alpes, LIP, and others. The "METADATA" section provides details like version 1.3.3, Software License (GNU Lesser General Public License v2.1 or later), Programming Language (C++), and Code Repository (<https://github.com/linbox-team/LinBox>). The "COLLECTION" section indicates it belongs to the LinBox collection. The "EXPORT" section offers options for CSV, XML, TBL, DC, and JSON formats. The "CITATION" section provides a BibTeX entry for the software.

This screenshot shows the Software Heritage archive page for the same identifier, [hal-02130801](https://hal.archives-ouvertes.fr/hal-02130801). The top navigation bar includes "Browse the archive", a search bar, and a "Permalink" link. The main content area shows the revision [e8e8328952266b7875c692963b11963b1496107](https://hal.archives-ouvertes.fr/e8e8328952266b7875c692963b11963b1496107), authored by Software Heritage on June 13, 2019, at 08:12 UTC. It was deposited in the "hal" collection. The "config-blas.h" file is displayed, showing its content. A large blue arrow points from the HAL page to this Software Heritage page, indicating the integration and availability of the same software through both platforms.

`swh:1:dir:393b611a1424f032e83569bf6762502371cfef65`

FAIRCORE4EOSC: Dagstuhl - SWH to implement same as [the HAL example above](#)

# Growing adoption of SWH in Academia (selection)

## HAL software curated deposit workflow

*Curated Archiving of Research Software Artifacts*

International Journal of Digital Curation, 2020

## 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<sup>A</sup>T<sub>E</sub>X class

## Policy: France



*National Plan for Open Science and Research Infrastructures*

## Policy: Europe



*EOSC SIRS report*

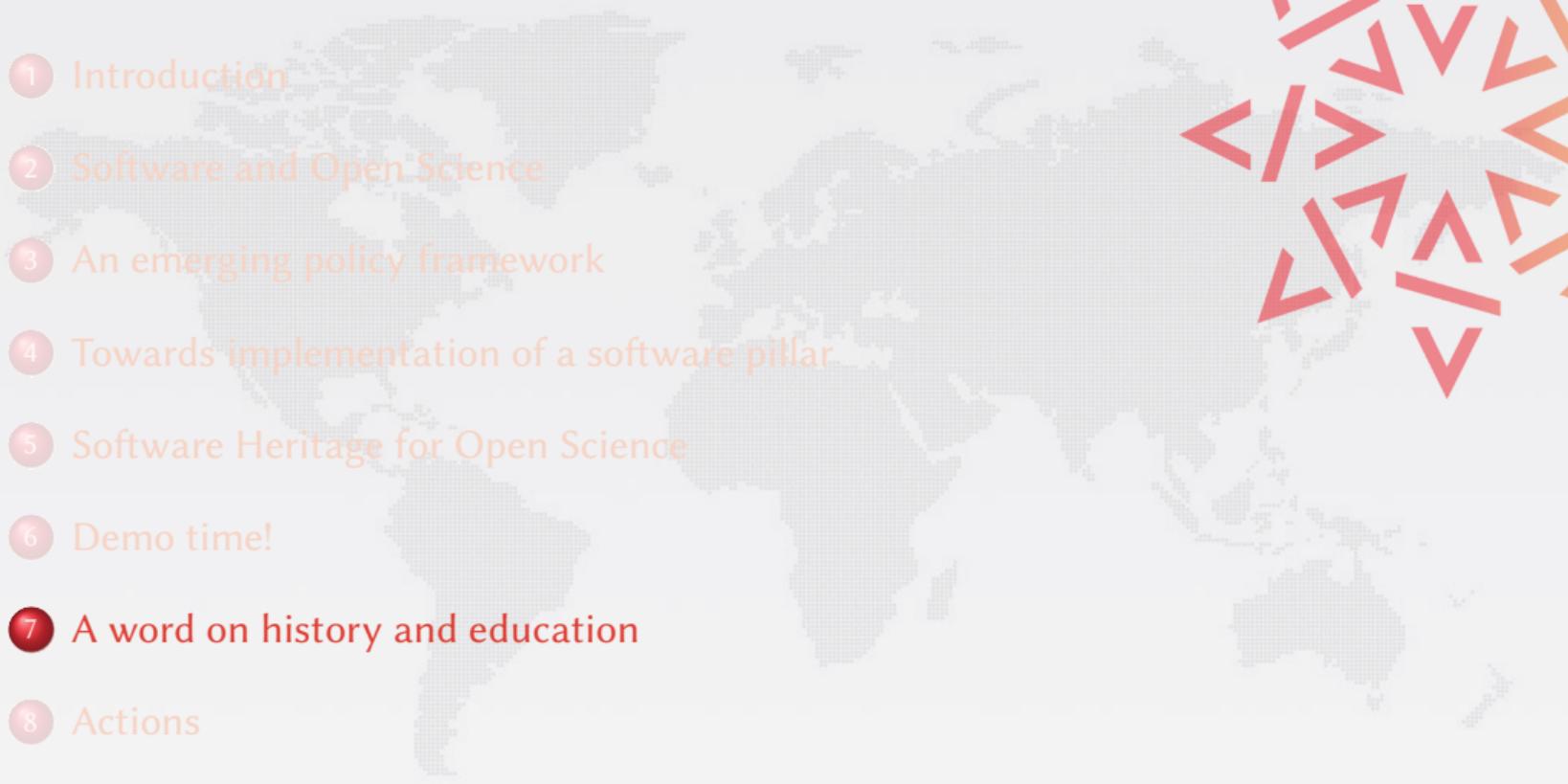
- SWHIDs
- archive

## Guidelines



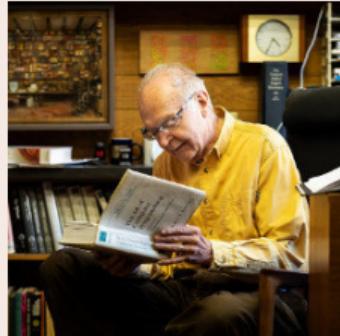
Software Heritage  
1 Prepare your public repository  
README, AUTHORS & LICENSE files  
2 Save your code  
<http://softwareheritage.org/>  
3 Reference your work  
(full repository, specific version or code fragment)

- summary
- ICMS 2020

- 
- 
- 1 Introduction
  - 2 Software and Open Science
  - 3 An emerging policy framework
  - 4 Towards implementation of a software pillar
  - 5 Software Heritage for Open Science
  - 6 Demo time!
  - 7 A word on history and education
  - 8 Actions

# Calling for preservation: Donald Knuth and Len Shustek

Communications of the ACM, February 2021



*"Telling historical stories is the best way to teach. It's much easier to understand something if you know the threads it is connected to."*

*Let's Not Dumb Down the History of Computer Science*

Donald E. Knuth, Len Shustek

<https://doi.org/10.1145/3442377>

A unique opportunity

most of the creators are still here: we can talk to them!

but the clock is ticking...



## 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”



- **Detailed process for Legacy Software**

- *curation*
  - reconstruction of the development history
  - collecting metadata
- *archival*
  - in Software Heritage

- **Traceability** of all process phases

- using modern version control tools

## Paris Call on Software Source Code

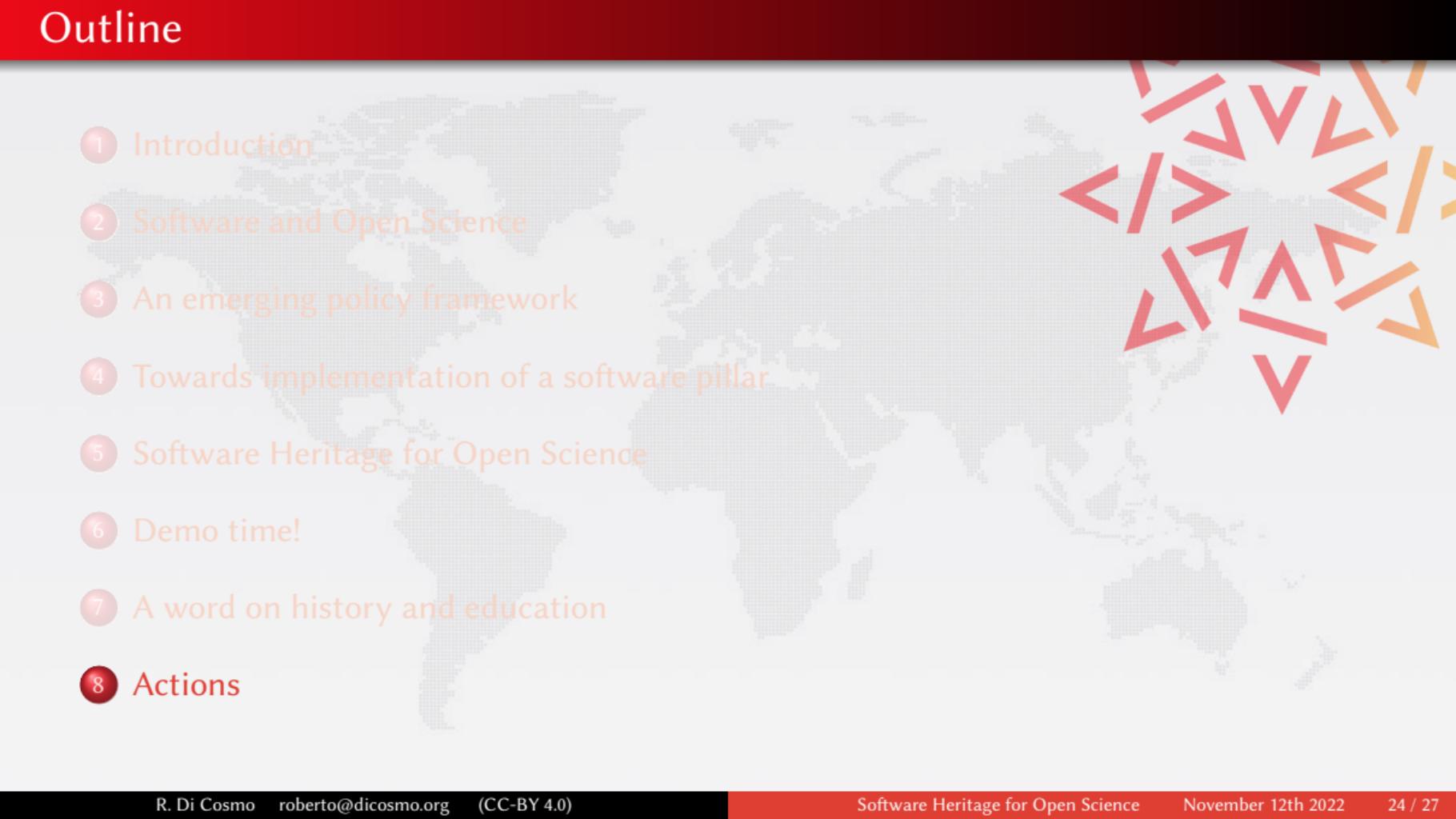
“[We call to] support efforts to gather and preserve the artifacts **and narratives** of the history of computing, while the earlier creators are still alive”



- **Expand** the SWHAP scope to
  - documents
  - media (videos, pictures, images, etc.)
  - oral history
- **Preserve and Present** all this material
- **Share** process and tools (all open source!)
  - with museums, archives and all interested parties

see this live on [the Software Stories website](#), and get [the guide](#) and [the SWHAP Days hybrid event](#), 19 and 20 october 2022



- 
- 1 Introduction
  - 2 Software and Open Science
  - 3 An emerging policy framework
  - 4 Towards implementation of a software pillar
  - 5 Software Heritage for Open Science
  - 6 Demo time!
  - 7 A word on history and education
  - 8 Actions

# Call to action: best practices for ARDC are available... today!

## Archiving and referencing

For **all source code used in research (yes, even small scripts!)**

- archive and reference in Software Heritage (see [detailed HOWTO](#))

## Describing and Citing/Crediting

For **software one wants to put forward**, add these **extra steps**:

- add **codemeta.json** with description (see the [codemeta generator](#))
- (french partners) reference in the HAL portal (see [online HAL documentation](#))
- cite software using the [biblatex-software](#) package (in CTAN and TeXLive)

## We can (and must)

- train students and colleagues
- engage journals, conferences, learned societies

# Policy remarks on the road ahead

Infrastructures for Software: avoid balkanisation, mutualise cost

- build on common, shared, open, non profit infrastructures
- join Software Heritage
  - development member/sponsor, mirror, contributor
  - adoption ambassador, learned societies, policy
  - research address the many scientific challenges

Walking the talk in Europe

ongoing full workpackage in [FAIRCORE4EOSC](#) interconnects infrastructures with Software Heritage

open now [CHIST-ERA joint ORD call](#) deadline: 14/12/2022

*Belgium, Czech Republic, France, Lithuania, Luxembourg, Poland, Slovakia, Switzerland, Turkey*

*"Processes and tools to describe, share, reference and archive software [...] that leverage existing initiatives, such as Software Heritage"*

# Policy remarks on the road ahead, cont'd

Avoid proprietarization of public research result

set the default to open *for software too*

- *publicly funded research software should be open source, exceptions must be justified*

Establish intelligent incentives

*When a measure becomes a target, it stops being a good measure.*

*Goodhart's law*

In hiring, funding and career evaluation

- avoid *purely numerical* indicators
- count *quality* software contributions in careers (all aspects!)
- keep *the human* in the loop

Remember to preserve the history

find, curate, archive and showcase landmark academic software **now...**

# Questions?

## References

-  UNESCO, *Draft recommendations on Open Science*  
2021, [\(online\)](#)
-  French Ministry of Research, *Second National Plan for Open Science*  
2021, [\(online\)](#)
-  EOSC SIRS Task Force, *Scholarly Infrastructures for Research Software*  
2020, Publications office of the European Commission, [\(10.2777/28598\)](#)
-  R. Di Cosmo, *Archiving and Referencing Source Code with Software Heritage*  
International Conference on Mathematical Software 2020 [\(10.1007/978-3-030-52200-1\\_36\)](#)
-  J.F. Abramatic, R. Di Cosmo, S. Zacchiroli, *Building the Universal Archive of Source Code*  
CACM, October 2018 [\(10.1145/3183558\)](#)