Towards a Software Pillar for Open Science
from policy to implementation

Roberto Di Cosmo
Chair, Software Chapter, National Committee for Open Science
Director, Software Heritage
Inria and Université de Paris Cité

September 9th 2022
Outline

1. Software and Open Science
2. An emerging policy framework
3. Towards implementation: assessing the needs for a software pillar
4. Conclusion
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.”

R. Di Cosmo  roberto@dicosmo.org (CC-BY 4.0)
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.”
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 Commissionner 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.
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 Commissionneer 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.”
Software is a pillar of Open Science

Software powers modern research

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

Top 100 papers (Nature, 2014)

Sometimes, if you don't have the software, you don't have the data

Christine Borgman, Paris, 2018
Software is a pillar of Open Science

Software powers modern research

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

*Top 100 papers (Nature, 2014)*

Sometimes, if you don't have the software, you don't have the data

*Christine Borgman, Paris, 2018*
Software is a pillar of Open Science

Software powers modern research

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

Sometimes, if you don’t have the software, you don’t 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 is a pillar of Open Science

Software powers modern research

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

*Top 100 papers (Nature, 2014)*

Sometimes, if you don’t have the software, you don’t 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
Software powers modern research

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

Sometimes, if you don’t have the software, you don’t have the data
Christine Borgman, Paris, 2018

Nota Bene

software may be a tool, a research outcome and a research object
access to the source code is essential!
Software is a pillar of Open Science

Software powers modern research

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

Top 100 papers (Nature, 2014)

Sometimes, if you don't have the software, you don't 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
“Programs must be written for people to read, and only incidentally for machines to execute.”

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

```
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    CODE508  # ASTRONAUT: PLEASE CRANK THE
          TC     BANKCALL  # SILLY THING AROUND
          CADR   GOPERF1
          TCF    GOT0060H  # 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
```

Len Shustek, Computer History Museum 2006

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

R. Di Cosmo roberto@dicosmo.org (CC-BY 4.0)
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 CODE508 # ASTRONAUT: PLEASE CRANK THE
    TC BANKCALL # SILLY THING AROUND
    CADR GOPERF1
    TCF GOTOP00H # 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 ...
```

Quake III source code (excerpt)

```
float Q_sqrt(float number)
{
    long i;
    float x2, y;
    const float threehalves = 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 * ( threehalves - ( x2 * y * y ) ); // 1st iteration
    // y = y * ( threehalves - ( 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 GOTOP00H # 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 threehalves = 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 * ( threehalves - (x2 * y * y ) ); // 1st iteration
    // y = y * ( threehalves - (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.”
Outline

1. Software and Open Science
2. An emerging policy framework
3. Towards implementation: assessing the needs for a software pillar
4. Conclusion
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”
**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"
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

SECOND FRENCH PLAN FOR OPEN SCIENCE
Generalising open science in France 2021-2024

Launch on 6 July 2021 by Frédérique Vidal, Minister for Higher Education, Research and Innovation

- Multiplying the levers for change in order to generalise open science practices
- Structuring the policy for opening up or sharing research data
- New commitments to the opening of source code produced by research
- European and international inclusion in the context of the French Presidency of the European Union
- Disciplinary and thematic variations: open science policies must be adapted to disciplinary specificities

Path Three: Opening up and promoting source code produced by research
- Recognize and support the dissemination under an open source license of software produced by publicly funded research programmes
- Highlight the production of source code from higher education, research and innovation
- Define and promote an open source software policy
- Produce a National Charter for Open Source Software coming from higher education, research and innovation
- Develop the link between data and software through a network of Chief Data Officers in the various universities and research performing organisations.
- Develop the economic models of open source software and make them known within commercialization services
- Support Software Heritage and recommend it for the archiving and referencing of source code
- Create an open source research software prize
- Provide greater recognition for software production in the career of researchers, research support staff
- Recognise source code as a contribution to research
- Develop proper coordination between software forges, open publication archives, data repositories and the scientific publishing sector.

Build an ecosystem that connects code, data and publications

« 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. »

Towards a Software Pillar of Open Science
R. Di Cosmo roberto@dicosmo.org (CC-BY 4.0)
September 9th 2022
SECOND FRENCH PLAN FOR OPEN SCIENCE
Generalising open science in France 2021-2024

Launch on 6 July 2021 by Frédérique Vidal, Minister for Higher Education, Research and Innovation

- Multiplying the levers for change in order to generalise open science practices
- Structuring the policy for opening up or sharing research data
- New commitments to the opening of source code produced by research
- European and international inclusion in the context of the French Presidency of the European Union
- Disciplinary and thematic variations: open science policies must be adapted to disciplinary specificities

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

1. Recognize and support the dissemination under an open source license of software produced by publicly funded research programmes
2. Highlight the production of source code from higher education, research and innovation
3. Define and promote an open source software policy
4. Produce a National Charter for Open Source Software coming from higher education, research and innovation
5. Develop the link between data and software through a network of Chief Data Officers in the various universities and research performing organisations.
6. Develop the economic models of open source software and make them known within commercialization services
7. Support Software Heritage and recommend it for the archiving and referencing of source code
8. Create an open source research software prize
9. Provide greater recognition for software production in the career of researchers, research support staff

"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."
Software Chapter in the CoSO

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
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, …
Software Chapter in the CoSO

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, …

The Open Science award for Open Source research software

See the official page at MESRI
Outline

1. Software and Open Science
2. An emerging policy framework
3. Towards implementation: assessing the needs for a software pillar
4. Conclusion

R. Di Cosmo  roberto@dicosmo.org  (CC-BY 4.0)
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 ARDC and infrastructures

R. Di Cosmo  roberto@dicosmo.org (CC-BY 4.0)
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
## What is at stake

<table>
<thead>
<tr>
<th>ARDC</th>
<th>Before ARDC</th>
<th>Beyond ARDC</th>
</tr>
</thead>
</table>
| - Archive for retrieval (*reproducibility*)  
- Reference for identification (*reproducibility*)  
- Describe for discovery and reuse  
- Cite/Credit for credit and evaluation | - **Development** practices and tools (VCS, build system, test suites, CI, …)  
- **Opening up** towards a community (documentation, organization, communication) | - **Policies** (dissemination, reuse, careers!)  
- **Sustainability** (legal, economic etc.)  
- Technology transfer  
- Advanced technologies and tools (quality, traceability, etc.) |

Need training, best practices

---

R. Di Cosmo  
roberto@dicosmo.org  
(CC-BY 4.0)
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 ARDC and infrastructures
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)
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!

R. Di Cosmo  roberto@dicosmo.org  (CC-BY 4.0)
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
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:
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!
Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all
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
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
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

Software Heritage

Cultural Heritage | Industry | Research | Public Administration
The largest software archive, a shared infrastructure

Software Heritage

- **Source files**: 12,538,666,608
- **Commits**: 2,654,066,174
- **Projects**: 181,249,577
- **Directories**: 10,342,140,231
- **Authors**: 48,778,458
- **Releases**: 33,580,610
An international, non profit initiative built for the long term

Sharing the vision

And many more ...

www.softwareheritage.org/support/testimonials
An international, non-profit initiative built for the long term

Sharing the vision

And many more ...

Donors, members, sponsors

Diamond sponsor

Platinum sponsors

Gold sponsors

Silver sponsors

Bronze sponsors

www.softwareheritage.org/support/testimonials

R. Di Cosmo  roberto@dicosmo.org  (CC-BY 4.0)
Addressing the four needs (see ICMS 2020 for details)

Archive (12B+ files, 170M+ projects)

Software Heritage Archive
Merkle DAG + blob storage
Loading & deduplication

tar
origins
snapshots
releases
revisions
directories
contents

dsc
dsc
dsc
hg
git
git
git
git
svn
svn
svn
tar
zip
software
origins

Forges
GitHub
lister
GitLab
lister
Debian
lister
PyPi
lister

Package repos

Scheduling
Listing (full/incremental)

Reference (20 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
so/f_tware citation
style
biblatex-so/f_tware, v 1.2-2 now on CTAN

R. Di Cosmo roberto@dicosmo.org (CC-BY 4.0)
Addressing the four needs (see ICMS 2020 for details)

Archive (12B+ files, 170M+ projects)

- save.softwareheritage.org
- deposit.softwareheritage.org
Addressing the four needs (see ICMS 2020 for details)

Archive (12B+ files, 170M+ projects)

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

Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs

Now supported in SPDX 2.2, Wikidata etc.
Addressing the four needs (see ICMS 2020 for details)

Archive (12B+ files, 170M+ projects)

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

Intrinsic, decentralised, cryptographically strong identifiers, SWHIDs

Now supported in SPDX 2.2, Wikidata etc.

Reference (20 billion SWHIDs)

- Intrinsic metadata from source code
- Contributed the Codemeta generator

Describe
Addressing the four needs (see ICMS 2020 for details)

Archive (12B+ files, 170M+ 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.

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

R. Di Cosmo  roberto@dicosmo.org  (CC-BY 4.0)
Outline

1 Software and Open Science

2 An emerging policy framework

3 Towards implementation: assessing the needs for a software pillar

4 Conclusion
Open Science is growing, and Software is part of it

A working agenda for the Software Pillar of Open Science

- avoid proprietarisation: set the default to open
  - *publicly funded research software should be open source*, exceptions must be justified
  - set up institutional support
  - build common knowledge base for technology transfer offices

- establish intelligent and effective incentives
  - count quality software contributions in careers, avoid purely numerical indicators, keep the human in the loop (mind Goodhart’s law)

- avoid balkanisation, support mutualised common infrastructures
  - build on common, shared, open, non-profit infrastructures, like Software Heritage
  - acknowledge the predominant human component of digital infrastructures
    - recurrent funding of their cost
    - proper evaluation of their service
References

- European Open Science Conference (OSEC)
  2022, (online)
- 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. Zacchirol, Building the Universal Archive of Source Code
  CACM, October 2018 (10.1145/3183558)