Towards a Software Pillar for Open Science
challenges and opportunities

Roberto Di Cosmo
Director, Software Heritage
Inria and Université de Paris Cité

June 2nd 2022
PNRIA
Outline

1. Open Science
2. Building the software pillar of Open Science: assessing the needs
3. Focus on ARDC and infrastructures
4. Demo time!
5. Focus on broader policy issues
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*.”

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.”

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 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.

Yuval Noah Harari (on COVID 19)

“The real antidote [to epidemic] is scientific knowledge *and* global cooperation.”
Free Software, AKA: Open Source, FOSS, FLOSS,…

Software that offers to *its users* the freedom to:
- **use** the software
- **study** and **adapt** the software
- **distribute** software copies
- **distribute** modified copies
Free Software, AKA: Open Source, FOSS, FLOSS,…

Software that offers to its users the freedom to:

- use the software
- study and adapt the software
- distribute software copies
- distribute modified copies

From the ripple in the early days (~1980’s) to a tidal wave
Open Source

Free Software, AKA: *Open Source, FOSS, FLOSS,…*

Software that offers to *its users* the freedom to:
- **use** the software
- **study** and **adapt** the software
- **distribute** software copies
- **distribute modified copies**

From the ripple in the early days (~1980’s) to a tidal wave

**Free Software has changed the way software (even proprietary!) is**
- developed
- tested
- deployed
- maintained
- marketed
- sold
- designed
- taught
- …

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)

Quake III source code (excerpt)

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)

<table>
<thead>
<tr>
<th>P635P0T3</th>
<th>CA</th>
<th>BIT6</th>
<th># IS THE LR ANTENNA IN POSITION 1 YET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EXTEND</td>
<td>RAN</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>EXTEND</td>
<td>BZF</td>
<td>P635P0T4</td>
</tr>
<tr>
<td></td>
<td>CAF</td>
<td>CODE500</td>
<td># ASTRONAUT: PLEASE CRANK THE</td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>BANKCALL</td>
<td># SILLY THING AROUND</td>
</tr>
<tr>
<td></td>
<td>CAFD</td>
<td>GOPERF1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TCF</td>
<td>GOTOPOOH</td>
<td># TERMINATE</td>
</tr>
<tr>
<td></td>
<td>TCF</td>
<td>P635P0T3</td>
<td># PROCEED SEE IF HE'S LYING</td>
</tr>
<tr>
<td>P635P0T4</td>
<td>TC</td>
<td>BANKCALL</td>
<td># ENTER INITIALIZE LANDING RADAR</td>
</tr>
<tr>
<td></td>
<td>CADR</td>
<td>SETPOS1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>POSTJUMP</td>
<td># OFF TO SEE THE WIZARD ...</td>
</tr>
<tr>
<td></td>
<td>CADR</td>
<td>BURNBABY</td>
<td></td>
</tr>
</tbody>
</table>
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  CODE5008    # 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_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;
}
```
“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  G0PERF1
         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 ...
```

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;
}
```

Len Shustek, Computer History Museum

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

R. Di Cosmo  roberto@dicosmo.org  (CC-BY 4.0)
UNESCO, Inria, Software Heritage invite
40 international experts to meet in Paris
Experts call for greater recognition of software source code as heritage for sustainable development

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

The call is published on Feb 2019
Experts call for greater recognition of software source code as heritage for sustainable development

6 November 2018

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

The call is published on Feb 2019

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

1. Open Science

2. Building the software pillar of Open Science: assessing the needs

3. Focus on ARDC and infrastructures

4. Demo time!

5. Focus on broader policy issues
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
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
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
(documents,
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.)

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
Outline

1. Open Science
2. Building the software pillar of Open Science: assessing the needs
3. Focus on ARDC and infrastructures
4. Demo time!
5. Focus on broader policy issues
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)

2019: big bad news keep coming in

- summer 2019: BitBucket announces Mercurial VCS sunset
- july 2020: BitBucket erases 250,000 repositories (including research software)
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)

2019: big bad news keep coming in

- summer 2019: BitBucket announces Mercurial VCS sunset
- july 2020: BitBucket erases 250,000 repositories (including research software)

2021: … in Academia too

- october 2021: Inria’s old gforge is unplugged
  - breaks the build chain of the OCaml package manager (Opam)
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)

**2019: big bad news keep coming in**
- summer 2019: BitBucket announces Mercurial VCS sunset
- july 2020: BitBucket erases 250,000 repositories (including research software)

**2021: … in Academia too**
- october 2021: Inria’s old gforge is unplugged
  - breaks the build chain of the OCaml package manager (Opam)

**Bottomline**
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)

### 2019: big bad news keep coming in

- summer 2019: BitBucket announces Mercurial VCS sunset
- july 2020: BitBucket erases 250,000 repositories (including research software)

### 2021: … in Academia too

- october 2021: Inria’s old gforge is unplugged
  - breaks the build chain of the OCaml package manager (Opam)

### Bottomline

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

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

- Forges
- GitHub
- GitLab
- Software Heritage Archive
- Merkle DAG + blob storage
- Loading & deduplication
- Debian source
- pypi source
- Package repos
- Listing (full/incremental)
- Scheduling
- Origin
- GitHub
- GitLab
- Debian
- PyPi
- PyPi source
- GitHub source
- Package loader
- Git
- Mercurial
- Debian source
- pypi source
- Git
- SVN
- Tar
- Zip
- Software origins
- Forges

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

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

- save.softwareheritage.org
- deposit.softwareheritage.org
Addressing the four ARDC 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.
Addressing the four ARDC 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

R. Di Cosmo  roberto@dicosmo.org  (CC-BY 4.0)
Addressing the four ARDC 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
1. Open Science

2. Building the software pillar of Open Science: assessing the needs

3. Focus on ARDC and infrastructures

4. Demo time!

5. Focus on broader policy issues
A walkthrough

- Browse the archive
- **Trigger archival** of your preferred software in a breeze
- Get and use SWHIDs (**full specification available online**)
- 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 journal: an article from IPOL
- Curated deposit in SWH via HAL, see for example: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, …
- Extracting all the software products for Inria, for CNRS, for LIRMM or for Rémi Gribonval using HalTools
Call to action on ARDC: let’s foster adoption!

Train students and colleagues to **archive and reference relevant source code**

- full details in the **ICMS 2020** article
- short operational **HOWTO online**
Call to action on ARDC: let’s foster adoption!

Train students and colleagues to **archive and reference relevant source code**
- full details in the **ICMS 2020** article
- short operational **HOWTO online**

Engage conferences, journals, learned societies to use Software Heritage and SWHIDs
*APIs for *save code now* and *deposit* are available to integrate with*
- Research Articles
- Artifact Evaluation Committees
- Badging initiatives
Call to action on ARDC: let’s foster adoption!

Train students and colleagues to archive and reference relevant source code
- full details in the ICMS 2020 article
- short operational HOWTO online

Engage conferences, journals, learned societies to use Software Heritage and SWHIDs
APIs for save code now and deposit are available to integrate with
- Research Articles
- Artifact Evaluation Committees
- Badging initiatives

Help grow and structure the community
- Promote the ambassador program
- Encourage our institutions to
  - include Software Heritage in their Open Science policy
  - become member/sponsor
  - build a Software Heritage mirror (see ENEA)
Outline

1. Open Science
2. Building the software pillar of Open Science: assessing the needs
3. Focus on ARDC and infrastructures
4. Demo time!
5. Focus on broader policy issues
Second French Plan for Open Science

GENERALISING OPEN SCIENCE IN FRANCE 2021-2024

R. Di Cosmo roberto@dicosmo.org (CC-BY 4.0)
French National plan for Open Science, 2021-2024

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
  - 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

Breaking news: the Software Chapter of the CoSO is live!
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

R. Di Cosmo  roberto@dicosmo.org  (CC-BY 4.0)
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, …
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
Twenty-four active members

Chairs: Roberto Di Cosmo and François Pellegrini

- Florent CHUFFART (Univ Grenoble Alpe)
- Mélanie CLÉMENT-FONTAINE (Univ Paris-Saclay - Versailles Saint-Quentin)
- Laurent COSTA (UMR 7041 ArScAn)
- Ludovic COURTÈS (Inria)
- Sébastien GÉRARD (Univ Paris-Saclay, CEA, List)
- Mathieu GIRAUD (CNRS, Univ Lille)
- Timothée GIRAUD (CNRS)
- Jean-Yves JEANNAS (Univ Lille, AFUL)
- Nicolas JULLIEN (IMT Atlantique)
- Daniel LE BERRE (Univ Artois, CNRS)
- Violaine LOUVET (CNRS / GRICAD - Univ Grenoble Alpes)
- Camille MAUMET (Inria, Univ Rennes, CNRS, Inserm)
- Clémentine MAURICE (CNRS)
- Grégory MIURA (Univ Bordeaux Montaigne)
- Raphaël MONAT (LIP6, Sorbonne Université)
- Patrick MOREAU (CNRS)
- Sophie RENAUDIN (AP-HP)
- Nicolas ROUGIER (Inria, Univ Bordeaux, CNRS)
- Filippo RUSCONI (CNRS-Univ Paris-Saclay)
- François SABOT (IRD)
- Sylvie TONDA-GOLDSTEIN (Inria)
- Samuel THIBAULT (Univ Bordeaux) (Univ Paris-Saclay)
The floor is yours

it’s a long road, but together we can make it

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)