

Free and Open Source Software and Academia: the data is in!

Roberto Di Cosmo

Director, Software Heritage
Inria and Université de Paris Cité

17 April 2024



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

- 1 Introduction
- 2 Software and Open Science
- 3 French national policy leads to hard data: national award
- 4 French national policy leads to hard data: national survey
- 5 French national policy leads to hard data: large scale analysis
- 6 Towards a global view

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

2010 *IRILL* 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 French national policy leads to hard data: national award
- 4 French national policy leads to hard data: national survey
- 5 French national policy leads to hard data: large scale analysis
- 6 Towards a global view

Software is a Pillar of Open Science: a selection

Paris Call on Software Source code (2019, UNESCO)

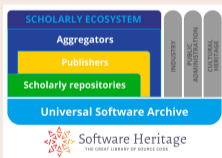


40 international experts call to “promote software development as a valuable research activity, and research software as a key enabler for Open Science/Open Research, [...] recognising in the careers of academics their contributions to high quality software development, in all their forms”



Open Source in UNESCO [recommendations](#) for Open Science, 2021

Software in the EOSC



2020 [EOSC SIRS](#) connect scholarly ecosystem via Software Heritage
2021 [EOSC Task Force](#) on Infrastructures for Research Software
2022 [FAIRCORE4EOSC project](#) WP6 implements SIRS report

And much more

Software track in [OSEC 2022](#), Software working group launched in Science Europe, DFG adds software [to model CV \(9/22\)](#), NASA unveils [Open Science policy \(12/22\)](#), ...

French National plan for Open Science, 2021-2024



SECOND FRENCH PLAN FOR OPEN SCIENCE

Generalising open science in France 2021-2024



1

Second French Plan for Open Science



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

2

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

7

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

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

8

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

9

Define and promote an open source software policy

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

3

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

Recognise source code as a contribution to research

- Create an **open source research software prize**
- **Provide greater recognition** for software production in the career of researchers, research support staff

Build an ecosystem that connects code, data and publications

- Develop **proper coordination** between software forges, open publication archives, data repositories and the scientific publishing sector.

4

- 1 Introduction
- 2 Software and Open Science
- 3 French national policy leads to hard data: national award**
- 4 French national policy leads to hard data: national survey
- 5 French national policy leads to hard data: large scale analysis
- 6 Towards a global view

National Open Science awards for FOSS in France: 2022 and 2023

First edition, 2022 prize



MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Accueil > Recherche > Science ouverte

Publié le 05.02.2022

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.

Sommaire

- The Coq proof assistant : lauréat de la catégorie Scientifique et technique
- Scikit-learn : lauréat de la catégorie Communauté
- Faust : lauréat de la catégorie Documentation
- Gammapy : prix du jury
- Jury

- 129 projects
- 4 awards
- 6 accessit
- first edition
- Coq proof assistant
- Scikit-Learn ML/AI
- Faust music
- Gammapy astronomy

Second edition, 2023 prize



MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Accueil > Recherche > Science ouverte

Publié le 29/11/2023

Remise des prix science ouverte du logiciel libre de la recherche 2023

Le ministère de l'Enseignement supérieur et de la Recherche remet pour la deuxième édition les Prix science ouverte du logiciel libre de la recherche. Huit logiciels développés par des équipes françaises sont récompensés pour leur contribution à l'avancée de la connaissance scientifique ou pour le caractère prometteur de leurs travaux.

Sommaire

- PfanGOOLIN : espoir de la catégorie « Scientifique et technique »
- Sireli : lauréat de la catégorie « Scientifique et technique »
- NoiseCapture : espoir de la catégorie « Communauté »
- OCami : lauréat de la catégorie « Communauté »
- KoOps : espoir de la catégorie « Documentation »
- Brian : lauréat de la catégorie « Documentation »
- Fink : espoir de la catégorie « Coup de cœur » du jury
- Hylio : lauréat de la catégorie « Coup de cœur » du jury
- Jury

- 66 projects
- 4 awards
- 4 "espoirs"
- will be run annually

Blueprint and data from the first edition (2021-2022)

Blueprints and analysis available



Open Research Europe

99 Views | 12 Downloads | 0 Citations

OPEN LETTER

Establishing a national research software award

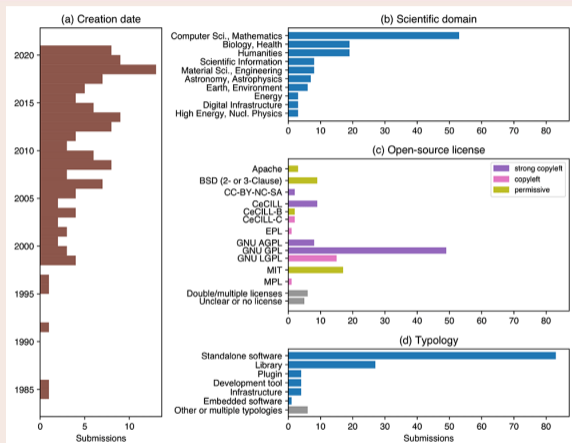
[version 1; peer review: 2 approved]

Isabelle Blanc Catala, Roberto Di Cosmo, Mathieu Graud, Daniel Le Berre, Violaine Louvet, Sophie Renaudin

College of experts for source code and software Committee for Open Science

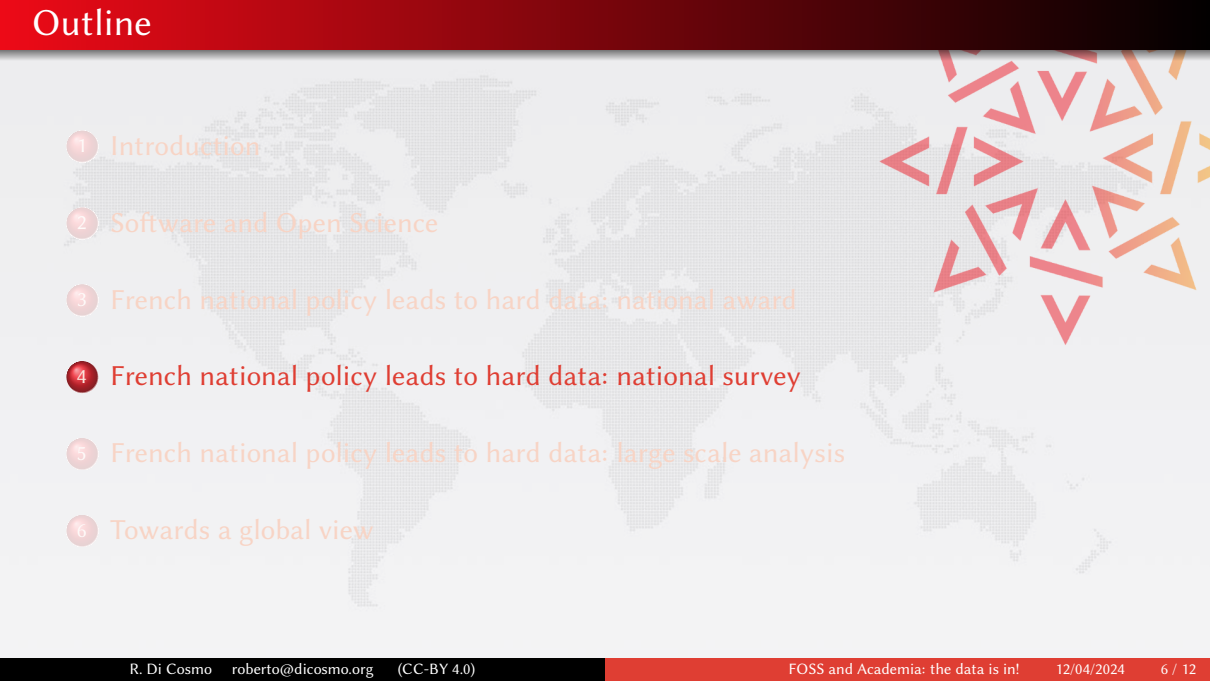
- goals, design decisions
- challenges and solutions
- lessons learned
- detailed data

A look at the data



Emulation is working

Australia: [here](#) and [here](#); Germany: [Helmholtz](#); European Commission: [a CSA](#); ...

- 
- 1 Introduction
 - 2 Software and Open Science
 - 3 French national policy leads to hard data: national award
 - 4 French national policy leads to hard data: national survey**
 - 5 French national policy leads to hard data: large scale analysis
 - 6 Towards a global view

National survey of research software (2023)

Context

Article 163 of Law No. 2022-217 of February 21, 2022 required a report on the production and impact of software resulting from research performed in publicly funded entities (universities, research organisations, etc.)

Process and selected results



Open survey (1331 detailed answers), and in depth exchanges with tech transfer offices

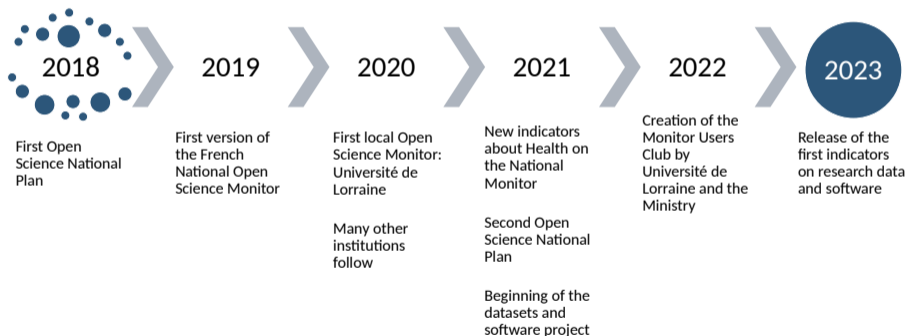
- 50% of software is older than 9 years
- 36% has more than 100 users
- 62% impact outside academia
- majority is FOSS, 10% proprietary
- 23% undergoes tech transfer

all details (in french) at



- 1 Introduction
- 2 Software and Open Science
- 3 French national policy leads to hard data: national award
- 4 French national policy leads to hard data: national survey
- 5 French national policy leads to hard data: large scale analysis
- 6 Towards a global view

A LITTLE BIT OF CONTEXT IN FRANCE...



Credits: Laetitia Bracco and the BSO team

MINING FULL-TEXTS TO DETECT MENTIONS TO DATASETS AND SOFTWARE

- **Innovative approach** based upon the use and development of machine learning tools
 - GROBID: full-text structuring
 - Softcite: **software mention detection**
 - DataStet: **data set mention detection**
- Automatic characterisation of mentions: **usage / production or creation / sharing**
- Another challenge: **downloading massive amounts of full-texts**



Alignments were carried out by [ClustalW](#) with default parameters (Thompson et al., 1994). The phylogenetic tree for the SIDREB gene was built using the software program [MEGA4](#) based on protein sequences. The phylogenetic tree was set up with the distance matrix using the Neighbor-Joining (NJ) method with 1000 bootstrap replications. Secondary structure prediction of the SIDREB protein was performed using the program [PSIPRED](#) (Jones, 1999). The *in vivo* structure prediction of the protein was done with the help of [I-TASSER](#) (Zhang, 2009). Automated homology model building of the DNA-binding domain was performed using the protein structure modelling program [MODELER](#) which models protein tertiary structure by satisfaction of spatial restraints. The input for [MODELER](#) consisted of the aligned sequences of IgG and the SIDREB, ensuring the that gives all the necessary commands to the [MODELER](#) to produce a homology model of the target on the basis of its alignment with the template. Energy minimization was performed by the steepest descent followed by the conjugate gradient method using a 20 Å non-bonded cut-off and a constant dielectric of 1.0. Evaluation of the predicted model involved analysis of the geometry and the stereochemistry of the model. The reliability of the model structure was tested using the [ENERGY](#) commands of [MODELER](#) (Zhang, 2009). The modelled structures were also validated using the program [PROSA](#) (Wiederstein and Sippl, 2007).

Southern blot analysis
Genomic DNA of *Sisymbrium* was extracted from leaves using the cetyltrimethylammonium bromide (CTAB) method (Saghai-Maroof et al., 1986), digested with *PvuII* and *BlnI* (New England Biolabs), fractionated in a 1.0% agarose gel, and blotted on a Hybond N⁺ membrane (Amersham). The blots were hybridized to a 705 bp SIDREB probe radioactively labeled with ³²P-dCTP using a High Prime DNA labeling kit (Roche, USA). Hybridization was carried out in 0.5 M sodium phosphate (pH 7.2), 7% SDS, and 1 mM EDTA.

Subcellular localization of the SIDREB protein
The SIDREB gene was fused to the 5' end of the green fluorescent protein (GFP) reporter gene using the pCAMBIA 1302 silent expression vector without a stop codon between the *NcoI* and *SpeI* sites. Recombinant DNA constructs encoding the SIDREB-GFP fusion protein downstream of the cauliflower mosaic virus (CaMV) 35S promoter were introduced into onion epidermal cells by gold particle bombardment using the PDS-1000 system (Bio-Rad) at 1100 psi helium pressure. Onion cells were also transiently transformed with the pCAMBIA 1302-GFP vector as a control. Transformed cells were placed on MS solid medium at 22°C and incubated for ~48 h before being examined. The subcellular localization of GFP fusion proteins was visualized with a confocal microscope (TCS, SP2, Leica).

I-TASSER

Type: software

Raw name: I-TASSER

References:

(Zhang, 2008) Zhang (2009) ^

authors Yang Zhang

title I-TASSER: Fully automated protein structure prediction in CASP8

date 2009

journal Proteins: Structure, Function, and Bioinformatics

volume 77

issue 59

first 100

page 113

last page 113

ISSN 0887-3585

DOI 10.1002/prot.22588

PMC ID PMC2782770

PMID 19786887

Open <http://neuro.oxfordjournals.org/doi/full/10.1093/bioinformatics/btp114>

Access pdf-render

publisher Wiley

I-TASSER (Iterative Threading ASSEMBLY Refinement) is a bioinformatics method for predicting three-dimensional structure model of protein molecules from amino acid sequences. It detects structure templates from the Protein Data Bank by a technique called



Credits: Laetitia Bracco and the BSO team

Uses improved version of SoftCite w.r.t. [the CZI 2022 study in biomedicine](#)

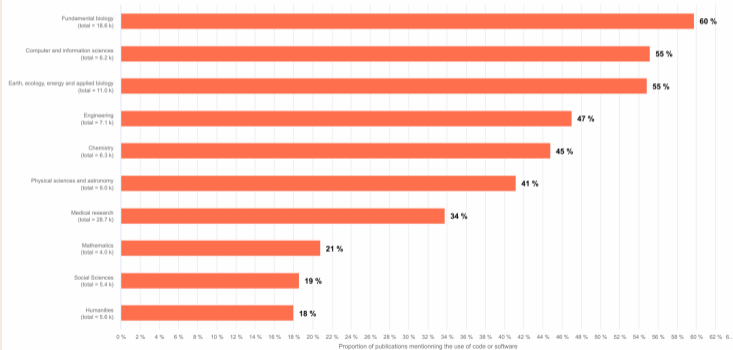
... gives a precious view, per discipline

Software Use

Proportion of publications in France that mention the use of code or software by discipline

Sort by:

Highest volume Highest use rate



Software is used massively across all disciplines/

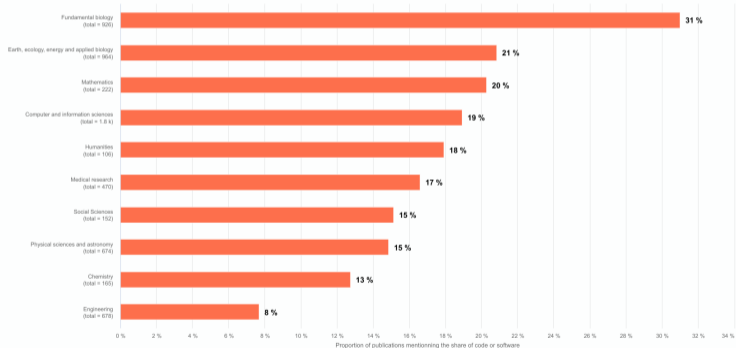


Software Sharing

Proportion of publications in France that mention code or software sharing by discipline

Sort by:

Highest volume Highest sharing rate



Over 20% of articles mentioning software creation actually **share it**



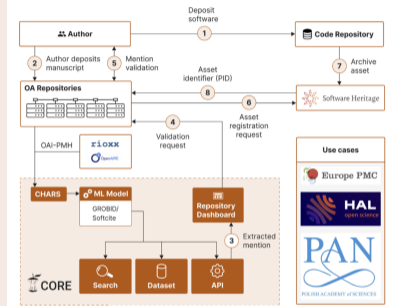
- 1 Introduction
- 2 Software and Open Science
- 3 French national policy leads to hard data: national award
- 4 French national policy leads to hard data: national survey
- 5 French national policy leads to hard data: large scale analysis
- 6 Towards a global view

Institutional: OSMF



Effort under **UNESCO** and **France impulse** to build an "open science monitor framework" compatible monitors across countries

Infrastructure: SOFair



Effort to identify software mentions in **all the open access literature**, and add links to the **Software Heritage archive**



FOSS and Open Science on a common path

- all research disciplines use software
- FOSS is the preferred approach for research software in France

The data is coming

massive analysis of research literature is now feasible

There is more than meets the eye, e.g.

- software projects classification
- advance software project search
- prior art identification
- ...

A common infrastructure



Software Heritage

Open non profit *universal source code archive* for industry, academia, public administration, culture and education

Questions?