

Logiciel, pilier de la science ouverte

défis et opportunités

Roberto Di Cosmo

Director, Software Heritage
Inria and Université de Paris Cité

Octobre 2024



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

- 1 Introduction
- 2 Logiciel et Code Source
- 3 La France montre la voie
- 4 Répondre aux besoins de l'ARDC
- 5 Indicateurs d'adoption et appel à l'action

Courte biographie : Roberto Di Cosmo

Professeur d'informatique à Paris, actuellement à INRIA

- 35+ *ans* de recherche (Informatique théorique, Programmation, Ingénierie logicielle, Numéro d'Erdos : 3)
- 25+ *ans* de Logiciel Libre et Open Source
- 15+ *ans* à construire et diriger des structures pour le bien commun



1999 *DemoLinux* – première distribution GNU/Linux live

2007 *Groupe thématique Logiciel Libre*
150 membres 40 projets 200Me

2008 *Projet Mancoosi* www.mancoosi.org

2010 *IRILL* www.irill.org

2015 *Software Heritage* à INRIA

2018 *Comité National pour la Science Ouverte*, France

2021 *EOSC Task Force sur les Infrastructures pour le Logiciel*,
Union Européenne

- 1 Introduction
- 2 Logiciel et Code Source
- 3 La France montre la voie
- 4 Répondre aux besoins de l'ARDC
- 5 Indicateurs d'adoption et appel à l'action

Harold Abelson, *Structure and Interpretation of Computer Programs* (1ère éd.)

1985

“Les programmes doivent être écrits pour être lus par des humains, et seulement accessoirement pour être exécutés par des machines.”

Code source d’Apollo 11 (extrait)

```
P63SP0T3      CA      BIT6      # IS THE LR ANTENNA IN POSITION 1 YET
EXTEND
RAND      CHAN33
EXTEND
BZF      P63SP0T4      # BRANCH IF ANTENNA ALREADY IN POSITION 1

CAF      CODE500      # ASTRONAUT:  PLEASE CRANK THE
TC      BANKCALL      #
CADR      GOPERF1
TCF      GOTOP00H      # TERMINATE
TCF      P63SP0T3      # PROCEED      SEE IF HE'S LYING

P63SP0T4      TC      BANKCALL      # ENTER      INITIALIZE LANDING RADAR
CADR      SETPOS1

TC      POSTJUMP      # OFF TO SEE THE WIZARD ...
CADR      BURNBABY
```

Code source de Quake III (extrait)

```
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, *Musée de l’Histoire de l’Informatique*

2006

“Le code source donne un aperçu de l’esprit du concepteur.”

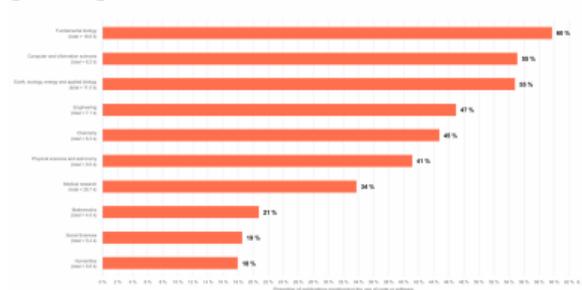
Le logiciel est un pilier de la Science Ouverte

Le logiciel alimente la recherche moderne

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

Sort by:

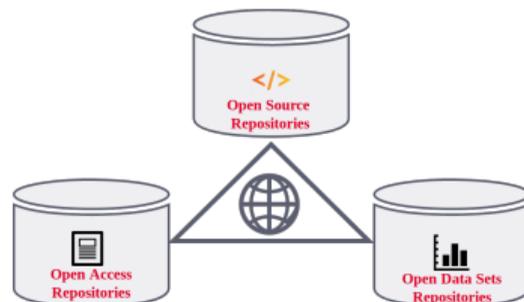
Highest volume Highest use ratio



Plus de 20% des articles utilisant des logiciels toutes disciplines confondues les partagent

2024 Observatoire Français de la Science Ouverte

Pilier clé : logiciel



Les liens sont **importants**

Nota Bene

le logiciel peut être un *outil*, un *résultat de recherche* et un *objet de recherche*

l'accès au *code source* est essentiel !

Préserver (l'historique du) code source est nécessaire pour la *reproductibilité*

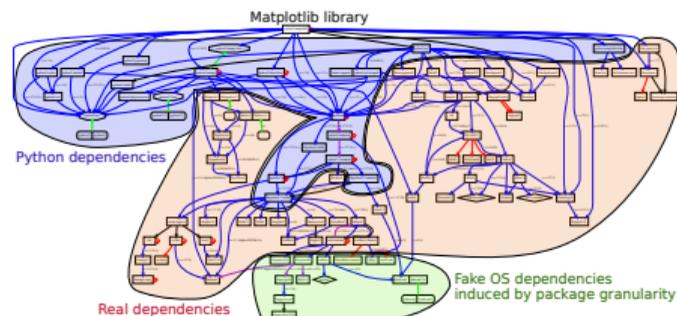
Le code source est *spécial* (le logiciel n'est *pas* une donnée)

Le logiciel *évolue* au fil du temps

- les projets peuvent durer des décennies
- l'*historique de développement* est la clé de sa *compréhension*

Complexité

- *millions* de lignes de code
- large *toile de dépendances*
 - facile à casser, difficile à maintenir
 - *logiciel de recherche* une fine couche supérieure
- communautés de développeurs *sophistiquées*



Le côté humain

conception, algorithmes, code, test, documentation, communauté, financement

et tant d'autres facettes ...

ARDC

- **Archiver** pour récupération (*reproductibilité*)
- **Référencer** pour identification (*reproductibilité*)
- **Décrire** pour découverte et réutilisation
- **Citer/Créditer** pour crédit et évaluation

Avant ARDC

- **Pratiques de développement** et outils (VCS, système de construction, suites de tests, CI, qualité du code, ...)
- **Ouverture** vers une communauté (documentation, organisation, communication)

Besoin de formation, d'outils, d'infrastructures, de bonnes pratiques

Au-delà de l'ARDC

- **Politiques** (diffusion, réutilisation, carrières, ...)
- **Durabilité** (juridique, financier, etc.)
- Transfert technologique
- Technologies et outils avancés (qualité, traçabilité, etc.)

- 1 Introduction
- 2 Logiciel et Code Source
- 3 La France montre la voie**
- 4 Répondre aux besoins de l'ARDC
- 5 Indicateurs d'adoption et appel à l'action

Plan national pour la science ouverte, 2021-2024

SECOND FRENCH PLAN FOR OPEN SCIENCE

Generalising open science in France 2021-2024



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

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

« 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

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

9

Define and promote an **open source software policy**

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.

Cinq axes d'action (voir détails en ligne)

- Identifier et mettre en valeur la production de logiciels de recherche
- Outils techniques et sociaux et meilleures pratiques
- Valorisation et durabilité
- Liaison et animation aux niveaux national, européen et international
- Reconnaissance et carrières

Introduction au Code Source concepts clés



- pour les étudiants
- pour les enseignants
- pour les chercheurs



Rapport sur les forges logicielles

dans le milieu
académique (FR) :

- besoins
- options
- limitations



Prix annuel

*Établir un prix national
pour les logiciels de
recherche. Open Research
Europe 2023*



Première édition, prix 2022



Accueil > Recherche > Science ouverte

Publié le 05.02.2022

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

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 projets
- 4 prix
- 6 accessits
- 1ère édition
- Coq assistant de preuve
- Scikit-Learn ML/IA
- Faust musique
- Gammapy astronomie

Deuxième édition, prix 2023



REUNION OFFICIELLE VOUS RESS

Accueil > Recherche > Science ouverte

Publié le 29/11/2023

Sommaire

- PParGGOLIN : espoir de la catégorie « Scientifique et technique »
- Smilei : lauréat de la catégorie « Scientifique et technique »
- NoiseCapture : espoir de la catégorie « Communauté »
- OCaml : lauréat de la catégorie « Communauté »
- KeOps : espoir de la catégorie « Documentation »
- Brian : lauréat de la catégorie « Documentation »
- Fink : espoir de la catégorie « Coup de cœur » du jury
- Hyphe : lauréat de la catégorie « Coup de cœur » du jury
- Jury

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.



- 66 projets
- 4 récompenses
- 4 "espoirs"
- désormais annuel

Plan et données de la première édition (2021-2022)

Plans et analyses disponibles



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 Giraud, Daniel Le Berre, Violaine Louvet, Sophie Renaudin.

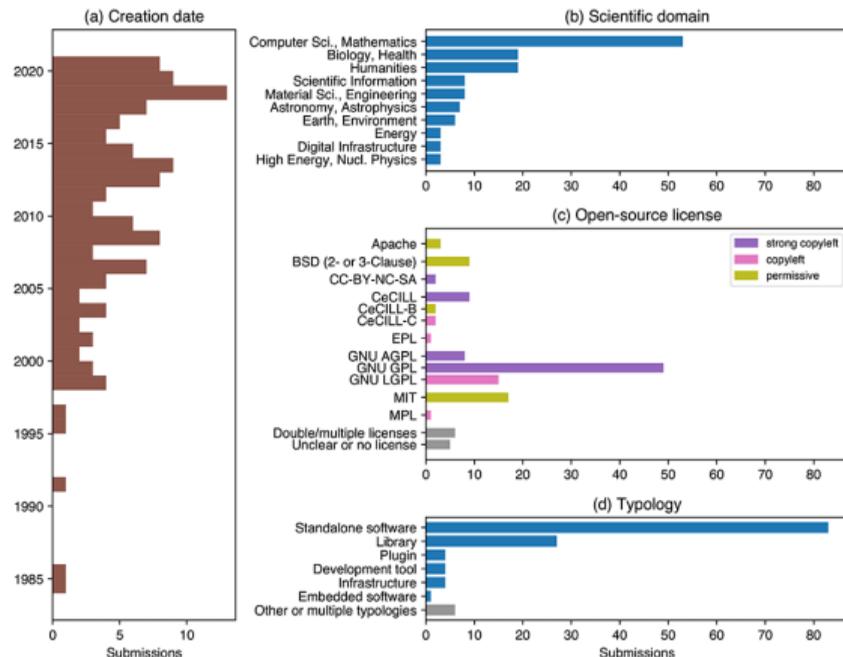
College of experts for source code and software Committee for Open Science

- objectifs, décisions de conception
- défis et solutions
- leçons apprises
- données détaillées

L'émulation fonctionne

Australie : [ici](#) et [ici](#) ; Allemagne : [Helmholtz](#) ; Commission européenne : [une CSA](#) ; ...

Un aperçu des données



Contexte

L'article 163 de la loi n° 2022-217 du 21 février 2022 exigeait un rapport sur la production et l'impact des logiciels issus de la recherche effectuée dans les entités financées par des fonds publics (universités, organismes de recherche, etc.)

Processus et résultats sélectionnés



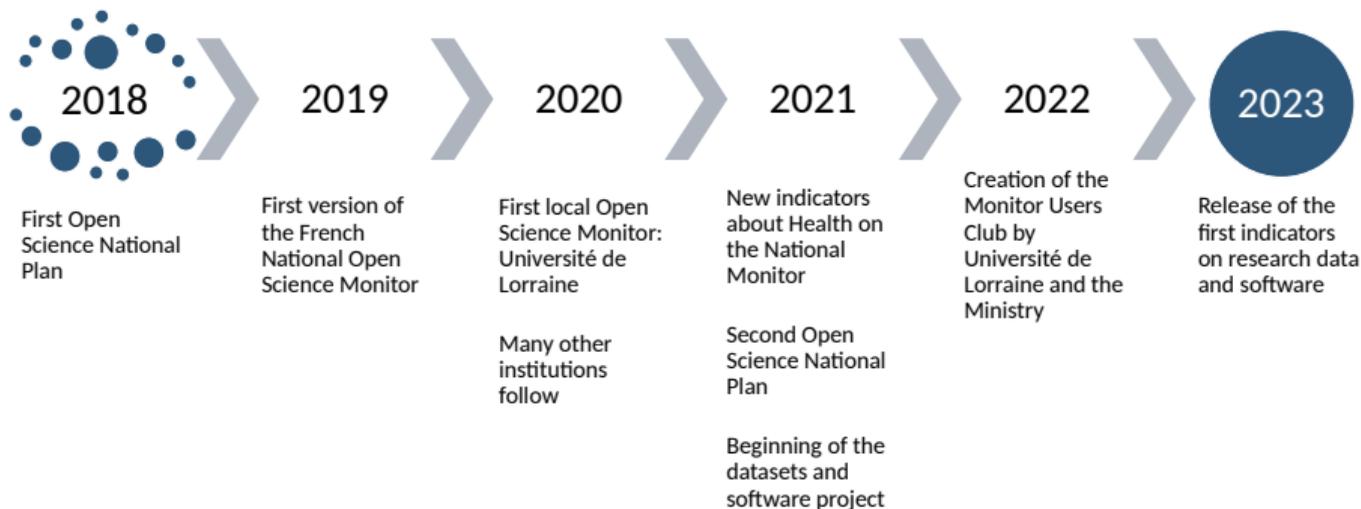
Enquête ouverte (1331 réponses détaillées), et échanges approfondis avec les bureaux de transfert de technologie

- 50% des logiciels ont plus de 9 ans
- 36% ont plus de 100 utilisateurs
- 62% ont un impact en dehors du milieu académique
- la majorité est en Logiciel Libre, 10% propriétaire
- 23% font l'objet d'un transfert technologique

tous les détails (en français) à



A LITTLE BIT OF CONTEXT IN FRANCE...



Crédits : Laetitia Bracco et l'équipe du BSO

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 [Castaño] with default parameters (Thompson et al., 1994). The phylogenetic tree for the SIDREB gene was built using the software program [MEGA] 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 ab initio structure prediction of the protein was done with the help of [I-TASSER] (Zhang, 2008). Automated homology model building of the DNA-binding domain was performed using the protein structure modeling program [MODELER] which models protein tertiary structure by satisfaction of spatial restraints. The input for [MODELER] consisted of the aligned sequences of Igt and the SIDREB protein, along with 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 produced 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 [PROCHECK-SL] (Moriya, 1992). The modeled structures were also validated using the program PROCRA (Wiederstein and Sippl, 2007).

Southern blot analysis
Genomic DNA of fossil millet was extracted from leaves using the cetyltrimethylammonium bromide (CTAB) method (Saghai-Maroof et al., 1984), digested with PvuII and HindIII (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 SIDREB² 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 SIDREB² gene was fused to the 5' end of the green fluorescent protein (GFP) reporter gene using the pCAMBIA 1302 plant 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 (2008) ^

authors Yang Zhang

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

date 2009

journal Proteins: Structure, Function, and Bioinformatics

volume 77

issue 99

first 100

page

last page 113

ISSN 0887-3585

DOI 10.1002/prot.22588

PMCID PMC2782770

PMID 19766687

Open <http://europepmc.org/articles/pmc27827>

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 et l'équipe du BSO

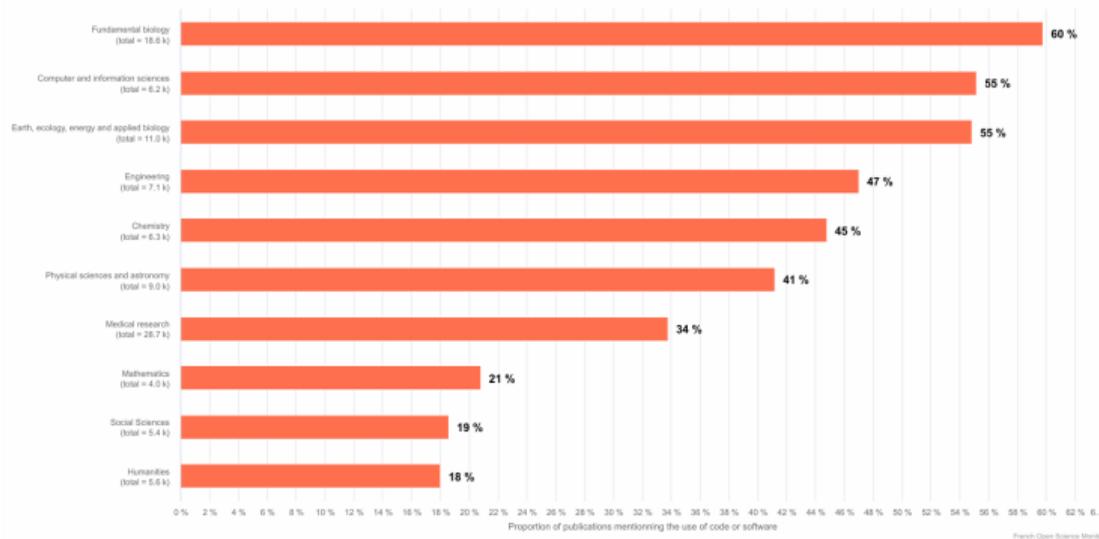
Utilise une version améliorée de SoftCite par rapport à l'étude CZI 2022 en biomédecine

Utilisation des logiciels

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

Sort by:

Highest volume Highest use rate



Les logiciels sont **massivement utilisés** dans toutes les disciplines

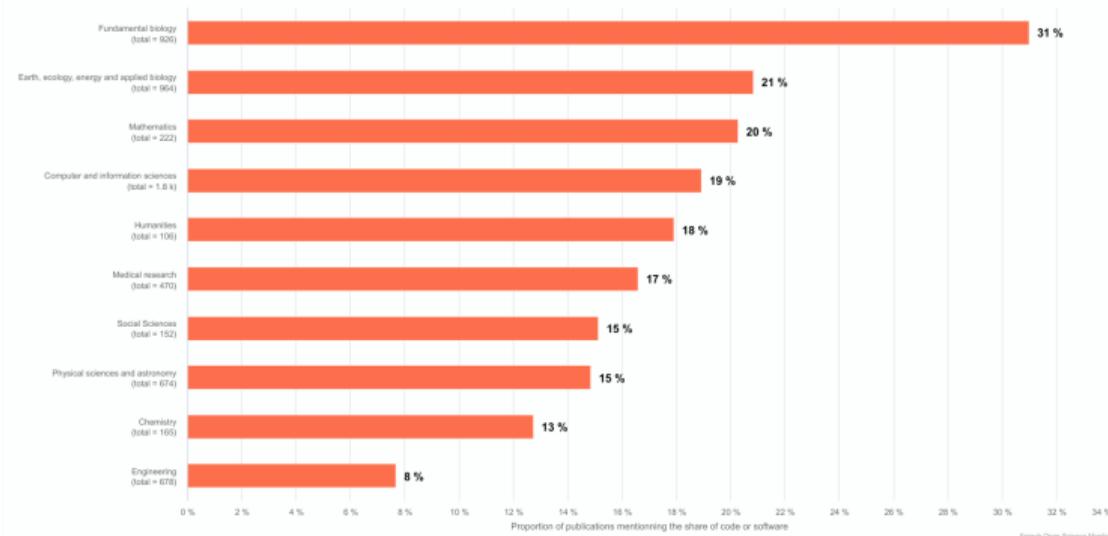


Partage des logiciels

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

Sort by:

Highest volume Highest sharing rate



Plus de 20 % des articles mentionnant la création de logiciels les **partagent réellement**



- 1 Introduction
- 2 Logiciel et Code Source
- 3 La France montre la voie
- 4 Répondre aux besoins de l'ARDC**
- 5 Indicateurs d'adoption et appel à l'action

Archivage

Les artefacts des logiciels de recherche doivent être correctement **archivés**
s'assurer que nous pouvons les *retrouver* (*reproductibilité*)

Référencement

Les artefacts des logiciels de recherche doivent être correctement **référencés**
s'assurer que nous pouvons les *identifier* (*reproductibilité*)

Description

Les artefacts des logiciels de recherche doivent être correctement **décrits**
faciliter leur *découverte* et leur *réutilisation* (*visibilité*)

Citer/Créditer

Les artefacts des logiciels de recherche doivent être correctement **cités** (*ce n'est pas la même chose que référencer !*)

donner du *crédit* aux auteurs (*évaluation!*)

Software Heritage : *un seul archive logiciel, une infrastructure partagée ...*

Une infrastructure ouverte et partagée



Inria



Le plus grand archive jamais construit



Diamond sponsor



Platinum sponsors



intel



Microsoft



Gold sponsors



Hugging Face

openinvention.ch

servicenow



Silver sponsors

AdaCore



GitHub

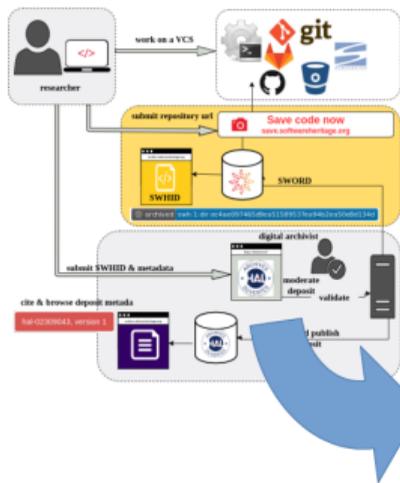
Google



Bronze sponsors



... meilleure expérience pour les chercheurs, dans le monde entier



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

The screenshot shows the HAL website for LinBox. The page title is 'Free and accessible knowledge'. The URL is <https://hal.archives-ouvertes.fr/hal-02130801>. The page content includes the LinBox logo, a list of contributing institutions (e.g., IRCC - Exact Computing, SWHID - Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier), and an abstract describing LinBox as a C++ template library for solving linear algebra problems. The page also features a 'BROWSE' section with a search bar and a list of related software packages.

The screenshot shows the Software Heritage archive page for LinBox. The page title is 'Browse the archive'. The URL is <https://hal.archives-ouvertes.fr/hal-02130801>. The page content includes the revision number 'e818328952266b775c692963b11963b1496107', the date '14 June 2019, 13:43 UTC', and the programming language 'C++'. The page also features a 'config-blas.h' file with its source code, which includes copyright information and license details.

sw.h1.dir:393b611a1424f032e83569bf6762502371cfc65

En savoir plus sur <https://softwareheritage.org/> demo à la demande

- 1 Introduction
- 2 Logiciel et Code Source
- 3 La France montre la voie
- 4 Répondre aux besoins de l'ARDC
- 5 Indicateurs d'adoption et appel à l'action

A few adoption indicators



Policy



- [Recommendations in ANR 2023 guidelines \(p. 17\)](#)
- HAL+SWH in [the Open Science software booklet](#)

Users and collaborations



What are they “referencing”?

source	n	percentage
Not available	2868	46.22
GitHub	1151	18.55
software heritage	387	6.24
zenodo	142	2.29
r package	70	1.13
cran	56	0.90
r package version	54	0.87
gitlab	35	0.56

Graphics Replicability Stamp Initiative



b/Surf: Interactive Bézier Splines on Surface Meshes

Claudio Mancinelli, Giacomo Nazzaro, Fabio Pellacini, Enrico Puppo
IEEE Transactions on Visualization and Computer Graphics (TVCG)



Repository



Projects



FAIRCORE4EOSC
Core Components Supporting a FAIR EOSC

The CodeMeta Project



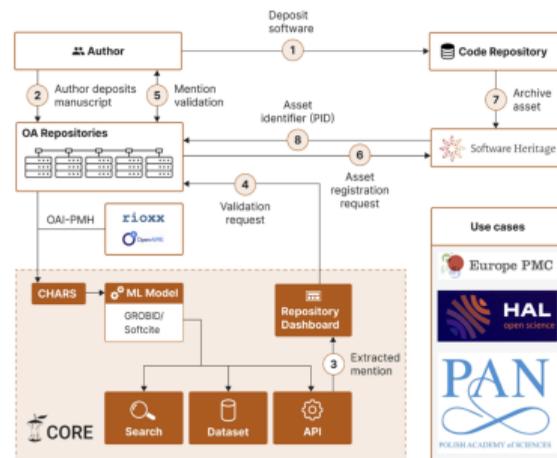
FAIR-IMPACT
Expanding FAIR solutions across EOSC

Institutionnel : OSMF



Effort sous l'impulsion de l'UNESCO et de la France pour construire un "cadre de suivi de la science ouverte" compatible avec les moniteurs à travers les pays

Infrastructure : SOFair



Effort pour identifier les mentions de logiciels dans toute la littérature en accès ouvert, et ajouter des liens vers l'archive Software Heritage



Construire un *catalogue uniforme* des logiciels de recherche

- métadonnées et PID pour le logiciel (utiliser CodeMeta et SWHID)
- point d'entrée unique entrer et extraire des informations (intégration HAL+SWH)
- informations sur tous les logiciels de recherche, ouverts **ou fermés**
- certaines informations sont privées (par exemple, détails de valorisation)

Nous avons

- HAL et SWH : dépôt modéré *pour le code ouvert avec métadonnées publiques*
- poussé à l'international (via EOSC, RDA, Force11)

Nous avons besoin (et les actions)

- **modérateurs des dépôts logiciel** dans chaque institution, formation via HAL/CCSD.
- étendre le catalogue pour couvrir *code fermé et informations privées*
- collaboration avec les équipes de valorisation

Comment participer

Contact : sabrina.granger@inria.fr

Appel à l'action : promouvoir les bonnes pratiques pour l'ARDC

Archiver et référencer

Tout **code source** utilisé dans la recherche (*même les petits scripts !*)

- sauver dans Software Heritage
- ajouter le **SWHID** dans les articles

Voir [guide détaillé en ligne](#)

reproductibilité



Décrire et Citer/Créditer

Pour **les logiciels que l'on souhaite mettre en avant** (*CV, rapports, etc., obtenir citations et du crédit*), suivre les **étapes supplémentaires** suivantes :

tutoriels vidéo

- ajouter un **codemeta.json** (voir le [générateur codemeta](#))
- référencer dans HAL ([documentation en ligne HAL](#))
- citer avec [biblatex-software](#) (dans CTAN et TeXLive)



former les étudiants et collègues

impliquer les revues, conférences, sociétés savantes

rejoindre le groupe d'intérêt Software Heritage ALIG



Développer une stratégie pour aborder l'ensemble des questions

- construire un corpus de connaissances partagées
 - travaux en cours aux "Ateliers de la donnée"
- construire un réseau d'expertise
 - se connecter avec d'autres institutions
 - se connecter avec des experts du logiciel libre et des OSPO
- développer un arbre de décision pour les chercheurs
- inclure le logiciel dans la politique Science Ouverte d'établissement

Comment procéder

- le réseau des ADAC
- les ateliers de la donnée
- le Collège logiciel dans le CoSO

travaillons ensemble pour mener la vague

Références

-  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](#))