Software Heritage

Infrastructure for Open Science

Roberto Di Cosmo
Director
roberto@dicosmo.org

March 24, 2026

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

Software is all around us

Invisible fabric of digital society

Knowledge is in the source

```
/**
 * @brief The basic unit of the simulation and is associated to a geographical location.
 *
 * Interventions (e.g., school closures) are tracked at this level. It contains a list of its
 * members (people), places (schools, universities, workplaces etc.), road networks, links to
 * airports etc.
 */
struct Microcell
{
        /* Note use of short int here limits max run time to USHRT_MAX*ModelTimeStep - e.g. 65536*0.25=16384 days=
           Global search and replace of 'unsigned short int' with 'int' would remove this limit, but use more memo
        */

        int n; // Number of people in microcell
        int adunit; // admin unit microcell belongs to
        int* members; // array of members/hosts of microcell

        int* places[MAX_NUM_PLACE_TYPES]; // list of places (of various place types) within microcell
        unsigned short int NumPlacesByType[MAX_NUM_PLACE_TYPES]; // number of places (of various place types) with
        unsigned short int keyworkerproph, move_trig, place_trig, socdist_trig, keyworkerproph_trig;
        unsigned short int move_start_time, move_end_time;
        unsigned short int place_end_time, socdist_end_time, keyworkerproph_end_time;
        TreatStat moverest, treat, vacc, socdist, placeclose;
        unsigned short int treat_trig, vacc_trig;
        unsigned short int treat_start_time, treat_end_time;
        unsigned short int vacc_start_time;
        IndexList* AirportList;
};
```

Covid Sim (excerpt)

Len Shustek, Computer History Museum 2006

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

A Global Undertaking

From all continents

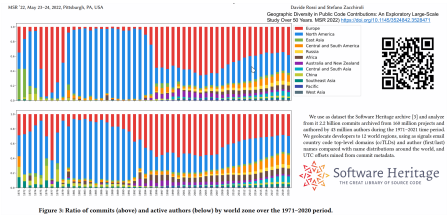
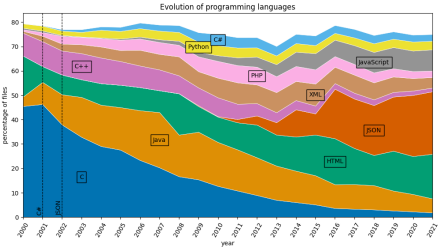
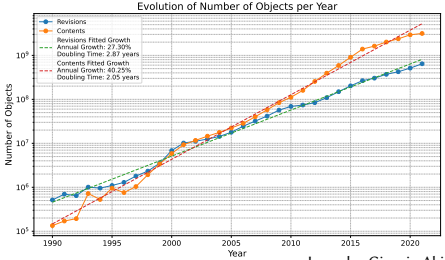


Figure 3: Ratio of commits (above) and active authors (below) by world zone over the 1971–2020 period.

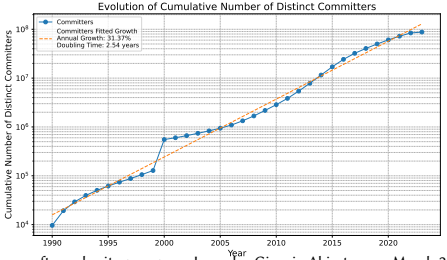
Many programming languages



An exponential growth (code)

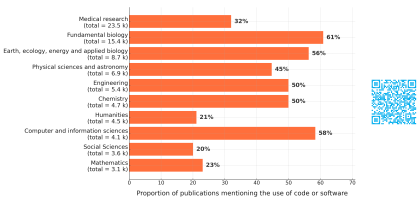


An exponential growth (contributors)



A pillar of Open Science

Pillar of Science across all research areas



A plurality of needs

Researcher

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

Laboratory/team track software contributions

- produce reports / web page

Research Organization know its software assets

- technology transfer
- impact metrics

we need a dedicated infrastructure: now we have it!

Outline

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Inria with unesco

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

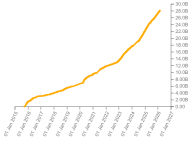
Inria with unesco

The largest open source code archive: one infrastructure, open, shared, non profit

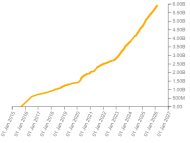
Unique digital common good *built in France since 2015*

Cultural Heritage | **Industry** | **Research** | **Public Administration**

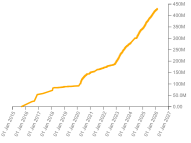
Source files
28,152,651,759



Commits
5,920,787,012



Projects
429,963,770



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Inria with unesco

The largest open source code archive: one infrastructure, open, shared, non profit

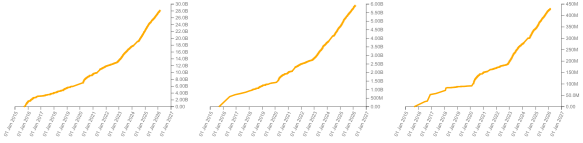
Unique digital common good *built in France since 2015*

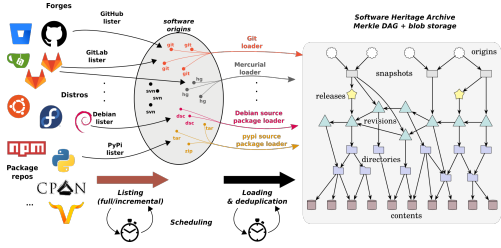
Cultural Heritage | **Industry** | **Research** | **Public Administration**

Source files
28,152,651,759

Commits
5,920,787,012

Projects
429,963,770





5000+ platforms

All versions, all history development in a single graph

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

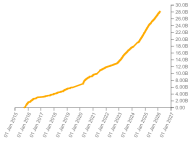
Inria with unesco

The largest open source code archive: one infrastructure, open, shared, non profit

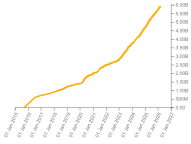
Unique digital common good *built in France since 2015*

Cultural Heritage | **Industry** | **Research** | **Public Administration**

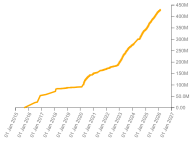
Source files
28,152,651,759

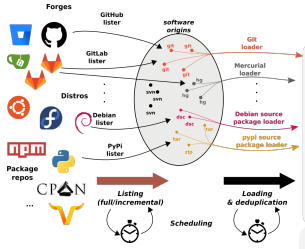


Commits
5,920,787,012



Projects
429,963,770





5000+ platforms

All versions, all history development in a single graph

- **50 × 10⁹** nodes
- **1000 × 10⁹** edges
- ~ **3 PB** of storage

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Inria with unesco

The largest open source code archive: one infrastructure, open, shared, non profit

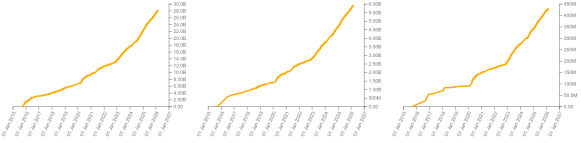
Unique digital common good *built in France since 2015*

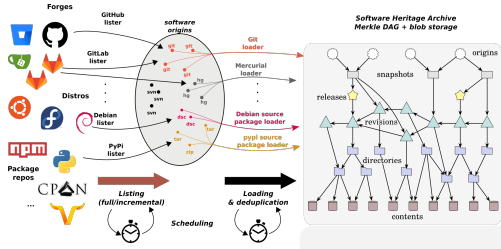
Cultural Heritage | **Industry** | **Research** | **Public Administration**

Source files
28,152,651,759

Commits
5,920,787,012

Projects
429,963,770





5000+ platforms

All versions, all history development in a single graph

- **50 × 10⁹** nodes
- **1000 × 10⁹** edges
- ~ **3 PB** of storage

A revolutionary **infrastructure** ensures **availability** guarantees **integrity** enables **traceability**

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

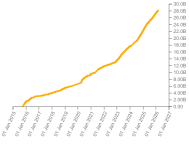
Inria with unesco

The largest open source code archive: one infrastructure, open, shared, non profit

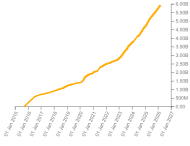
Unique digital common good *built in France since 2015*

Cultural Heritage | **Industry** | **Research** | **Public Administration**

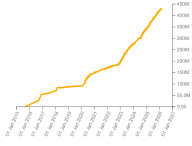
Source files
28,152,651,759

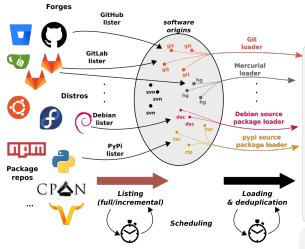


Commits
5,920,787,012



Projects
429,963,770





5000+ platforms

All versions, all history development in a single graph

- **50 × 10⁹** nodes
- **1000 × 10⁹** edges
- ~ **3 PB** of storage

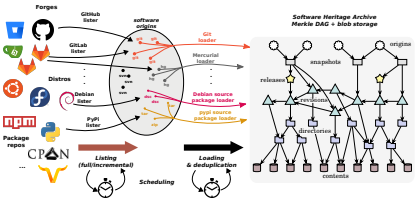
A revolutionary **infrastructure** ensures **availability** guarantees **integrity** enables **traceability**

IBM Microsoft TII Technology Innovation Institute HUAWEI cnrs Infosys cea Ministère de l'Enseignement Supérieur et de la Recherche aws Google Université Paris Cité openinventionnetwork

AdaCore GitHub Université di Pisa Sorbonne Université Université de Lorraine Sant'Anna NWO université PARIS-SACLAY

Addressing key needs in (Open) Science

Archive (28B+ files, 430M+ projects)



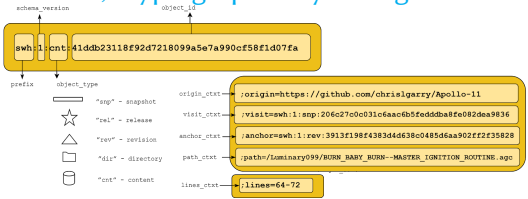
- save now, updateswh, webhooks
- deposit.softwareheritage.org

Describe

- *Intrinsic metadata* from source code
- Contributed the Codemeta generator

Reference (50 billion SWHIDs)

Intrinsic, cryptographically strong IDs



Now in SPDX 2.2, Wikidata

`https://swhid.org` - ISO/IEC 18670

Cite/Credit

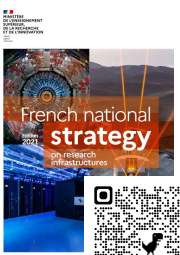
- Contributed biblatex-software style
- Software Citation from the archive!

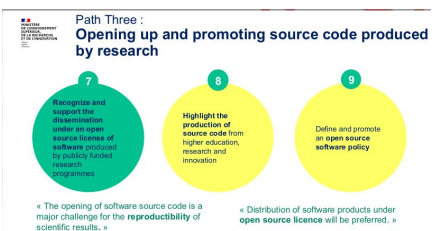
An example is worth a thousand words

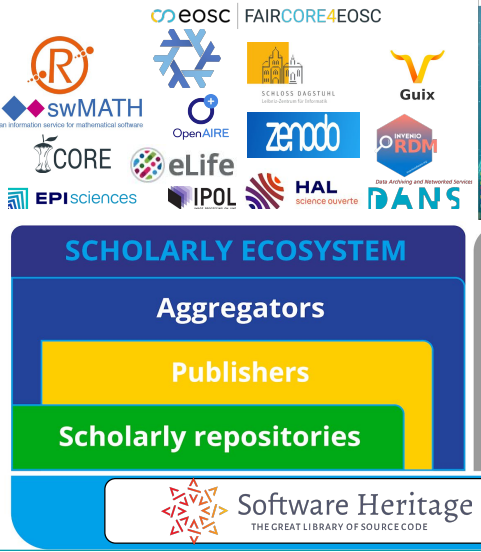
- Browse + Reference [ISO 18670] (Apollo 11 [excerpt], your work may be already there !)
- Trigger archival, use the updateswh browser extension, configure the webhooks
- Cite from the archive with biblatex-software (CTAN, ACMART)
- Describe with Codemeta (use codemeta generator)
- Curated deposit in SWH via HAL, see for example: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, ...
- Extracting all the software products for Inria, for CNRS, for CNES, for LIRMM or for Rémi Gribonval using HalTools
- Example with Parmap: devel on Github, archive in SWH, curated deposit in HAL
- Example research articles:
  - compare Fig. 1 and conclusions in the 2012 version and the updated version
  - SWHID in a replication experiment

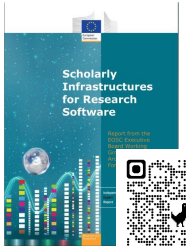
Outline

Connections and mutualization for the scholarly ecosystem





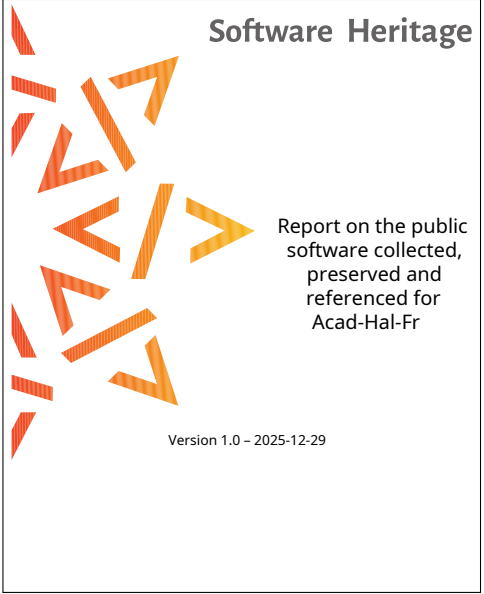




- [Funding agencies recommendations ANR 2023 guidelines (p. 17)]
- HAL+SWH in [the Open Science software booklet] from the French Ministry of Higher Education and Research

Outline

Forthcoming: Strategic Software Insights report



Contents

A peek inside the French Academic Software report

3531626

Contributions

46533

Software Projects

Active last 12 months: 3425

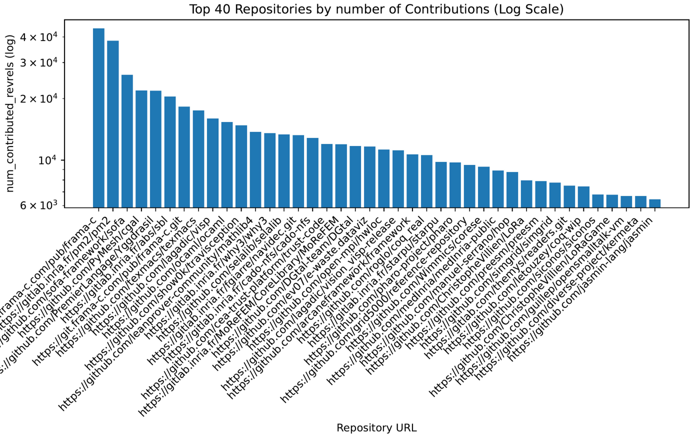
Still crawled: 44874

30101

Contributors

1992-07-08
2025-05-17

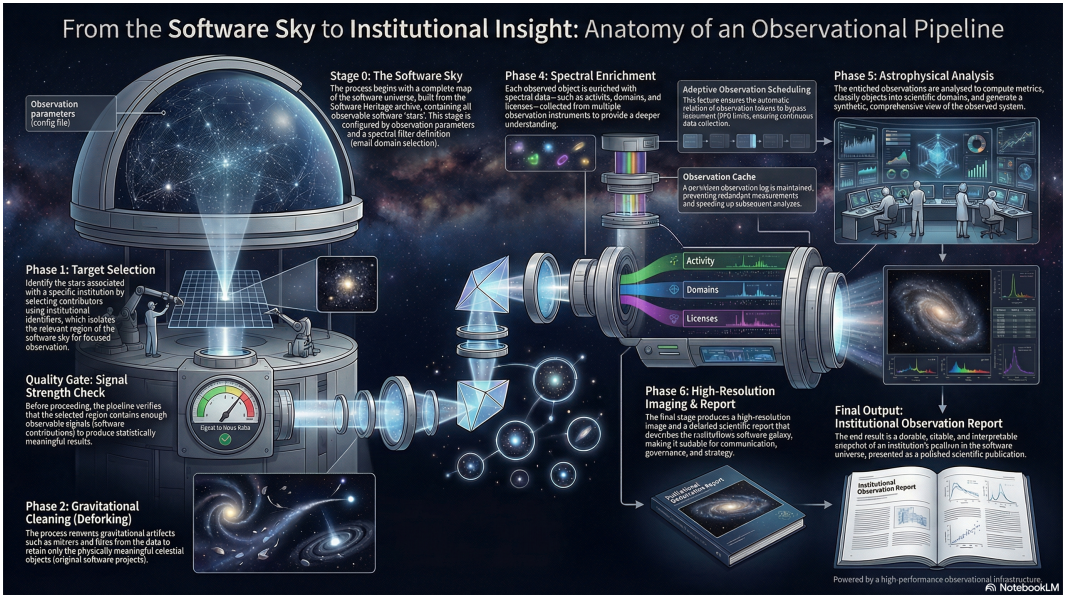
Time Span



domain	count
inria.fr	3833
univ-lyon1.fr	2806
univ-orleans.fr	1555
uca.fr	1371
jussieu.fr	1247
utc.fr	1216
cea.fr	1187
insa-toulouse.fr	747
ens-lyon.fr	706
univ-artois.fr	696
inrae.fr	690
univ-lemans.fr	688
cnrs.fr	682
ens.fr	675
inra.fr	644
univ-fcomte.fr	634
unilim.fr	541
insa-rennes.fr	491
univ-avignon.fr	489
univ-pau.fr	483
utt.fr	472
polytechnique.fr	437
uha.fr	422
pasteur.fr	420
telecom-paris.fr	412
univ-cotedazur.fr	363
univ-smb.fr	352
univ-tours.fr	350
u-picardie.fr	349
enpc.fr	328
sciencespo.fr	327
universite-paris-sa	323
univ-poitiers.fr	312
u-paris.fr	289
univ-angers.fr	267
univ-paris8.fr	264
inserm.fr	211
centralesupelec.fr	192
ird.fr	171
imt-atlantique.fr	159
onera.fr	158
univ-tlse2.fr	131
irstea.fr	130
enac.fr	126
univ-eiffel.fr	125
agroparistech.fr	105
cnam.fr	104
mnhn.fr	95
uphf.fr	94
univ-tln.fr	89

Note: Showing top 50 of 371 total domains

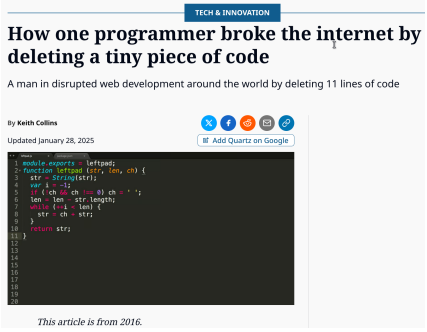
Software Heritage is the Very Large Telescope for source code



Outline

The next frontier: digital resilience

An early warning: left-pad (March 2016)



TECH & INNOVATION

How one programmer broke the internet by deleting a tiny piece of code

A man in disrupted web development around the world by deleting 11 lines of code

By Keith Collins

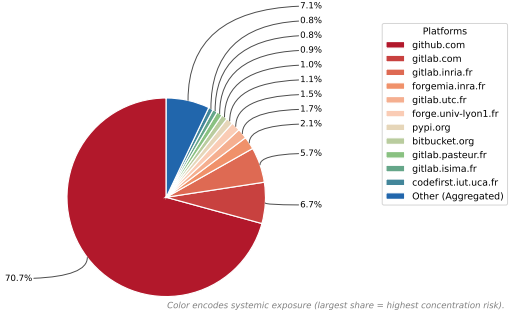
Updated January 28, 2025

Add Quartz on Google

```
module.exports = leftpad;
function leftpad (str, len, ch) {
  str = String(str);
  var i = -1;
  if (!ch && ch !== 0) ch = ' ';
  len = len - str.length;
  while (++i < len) {
    str = ch + str;
  }
  return str;
}
```

This article is from 2016.

A clear and present danger



Color encodes systemic exposure (largest share = highest concentration risk).

`French Academia` devs *(Source: Software Heritage)*

See the Wikipedia article:

- Facebook, PayPal, Netflix, Spotify affected
- web broken worldwide ~2.5 hours

"*If GitHub or PyPI disappeared tomorrow,*
all research would not just slow down
it would stop."

Guix and Nix integration show the way

Connecting reproducible deployment to a long-term source code archive

Ludovic Courtès — March 29, 2019

GNU Guix can be used as a “package manager” to install and upgrade software packages as is familiar to GNU/Linux users, or as an environment manager, but it can also provision containers or virtual machines, and manage the operating system running on your machine.

One foundation that sets it apart from other tools in these areas is *reproducibility*. From a high-level view, Guix allows users to *declare* complete software environments and instantiate them. They can share those environments with others, who can replicate them or adapt them to their needs. This aspect is key to reproducible computational experiments: scientists need to reproduce software environments before they can reproduce experimental results, and this is one of the things we are focusing on in the context of the Guix-HPC effort. At a lower level, the project, along with others in the Reproducible Builds community, is working to ensure that software build outputs are reproducible, bit for bit.

Work on reproducibility at all levels has been making great progress. Guix, for instance, allows you to travel back in time. That Guix can travel back in time *and* build software reproducibly is a great step forward. But there’s still an important piece that’s missing to make this viable: a stable source code archive. This is where Software Heritage (SWH for short) comes in.

When source code vanishes

When a platform or registry fails,

Guix fallbacks

to Software Heritage

and continues to build!

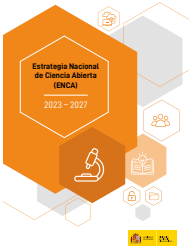
(*since 2019*)

Next step: scale up a sovereign fallback for Europe’s build chains

Outline

Software in the Open Science strategy

ENCA 2023-2027



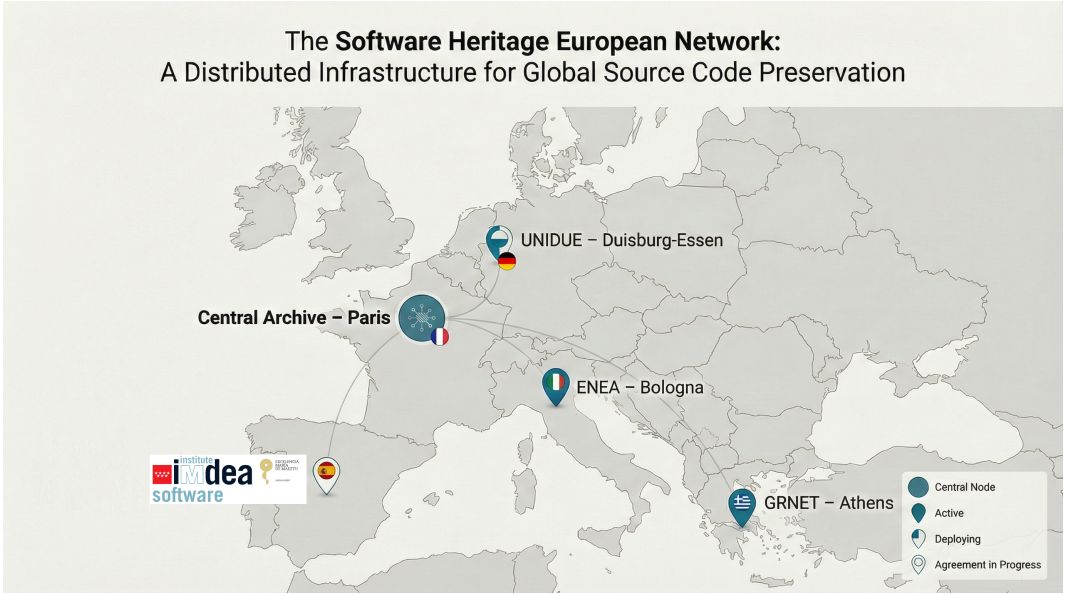
- Objective 1 (page 11): "[...] infraestructuras digitales interoperables suficientemente robustas y bien articuladas"
- Objective 2 (page 11): "Establecer nuevos mecanismos de evaluación de la investigación..."

French National Plan for Open Science 2021

Key Actions (selection)

- Define and promote an open source software policy
- Recognise source code as a contribution to research
- Create an open source research software prize
- Support Software Heritage for archiving and referencing source code

Software Heritage mirrors for *ecumenic digital sovereignty*



Food for thought

Foster adoption of best practices

archive, reference, describe, cite software: see the Software Heritage HOWTO

Establish incentives: funding, evaluation and recognition

- count *quality* software contributions (all aspects!)
- establish National/Regional/Institutional *awards* (get the blueprints from France)

Build objective indicators: *Strategic Insights Report, not surveys*

Avoid *balkanisation*, build upon shared, open, non profit infrastructures (POSI, UNESCO)

- FAIRCORE4EOSC interconnected infrastructures with Software Heritage
- OSPO-Radar institutional portals/reporting via Software Heritage outside France

Avoid proprietarization of public research result

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