

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Software Heritage face à l'IA : d'une archive ouverte à une infrastructure de recherche souveraine

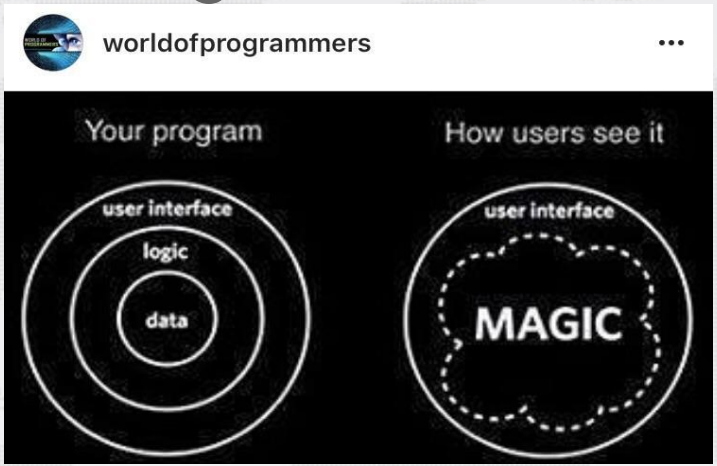
Morane Gruenpeter
Directrice des écosystèmes académiques
Software Heritage, Fondation Inria

Objectifs



- ★ Introduction : le savoir est dans le code source
- ★ Visite guidée : l'archive Software Heritage
- ★ La science ouverte pour la liberté du savoir
- ★ IA ouverte, responsable et transparente : un but commun
- ★ Vers la #SouverainetéNumérique

Qu'est-ce qu'un logiciel ?



Le logiciel en tant que concept

- **projet** ou entité
- la **communauté** autour du projet
- l'**idée** logicielle / algorithmes / solutions

Pas un artefact numérique

Artefacts logiciels

- Exécutables
  - Pour plusieurs environnements
- Code source

Une multitude d'artefacts numériques

Le savoir est dans le code source

« Le code source offre une fenêtre vers l'esprit de son concepteur. »

Len Shustek, 2006
Computer History Museum

Ordinateur de guidage d'Apollo 11 (~60 000 lignes), 1969

```
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     GOTOPOOH        # 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
```

Margaret Hamilton
(1936-2026)



Code source d'Apollo 11 dans SWH | Article

Conservé à la **bibliothèque de recherche du MIT**
numérisé par Rob Burkey depuis 2003

Code pour humains, binaires pour machines

« Les programmes doivent être écrits pour être lus par des humains, et seulement accessoirement pour être exécutés par des machines. » [1]

Manuel de 1984 ***Structure and Interpretation of Computer Programs*** **(SICP)**, écrit par **Harold Abelson et Gerald Jay Sussman**

Our design of this introductory computer-science subject reflects two major concerns. First, we want to establish the idea that a computer language is not just a way of getting a computer to perform operations but rather that it is a novel formal medium for expressing ideas about methodology. Thus, programs must be written for people to read, and only incidentally for machines to execute. Second, we believe that the essential material to be addressed by a subject at this level is not the syntax of particular programming-language constructs, nor clever algorithms for computing particular functions efficiently, nor even the mathematical analysis of algorithms and the foundations of computing, but rather the techniques used to control the intellectual complexity of large software systems.

Qui produit le savoir ?
Et qui le consomme ?

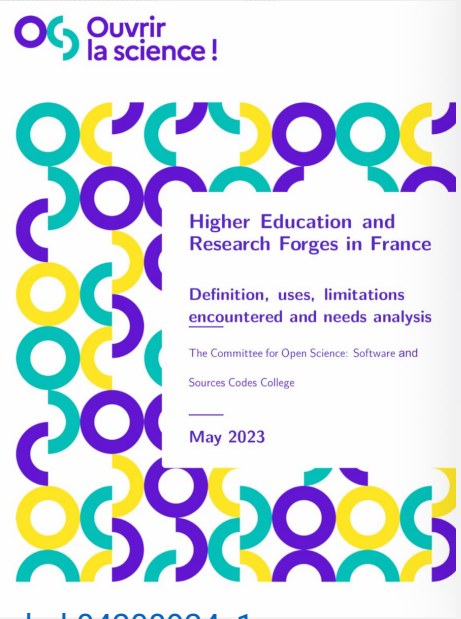
Nés et maintenu dans les forges

*« **Les forges logicielles** sont principalement conçues pour gérer les artefacts du génie logiciel (comme le code source), les fichiers binaires et même la documentation tout au long de leur cycle de vie. Les forges peuvent aussi servir à rédiger collaborativement des articles scientifiques ou de la documentation, voire à écrire des pages web à l'aide d'outils de traitement de fichiers texte peu structurés… »*





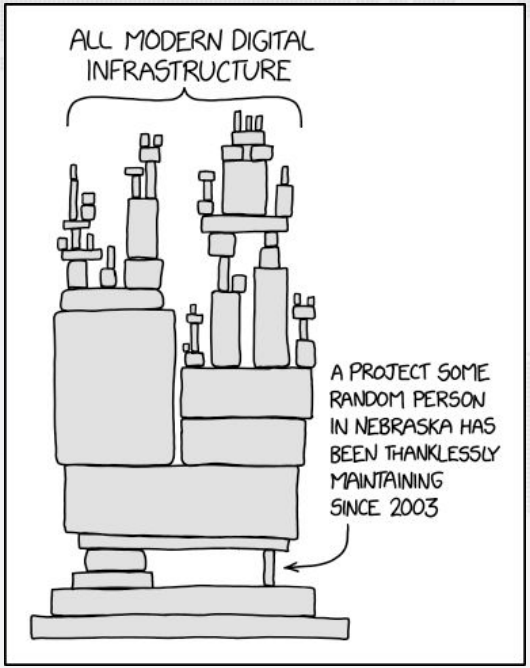
GitHub



hal-04208924v1

Le code source est…

Le socle invisible de l'infrastructure numérique



https://xkcd.com/2347/

La #SouverainetéNumérique est-elle possible sans le code ?

Le code source est fragile !



Google Kills Off Google Code

Natasha Lomas @riptari / 10:58 AM GMT+1 • March 13, 2015

1,4 M de projets

SD Times

Google begins shutdown of its code repository

Latest News Published: March 12th, 2015 - Michael Pehel

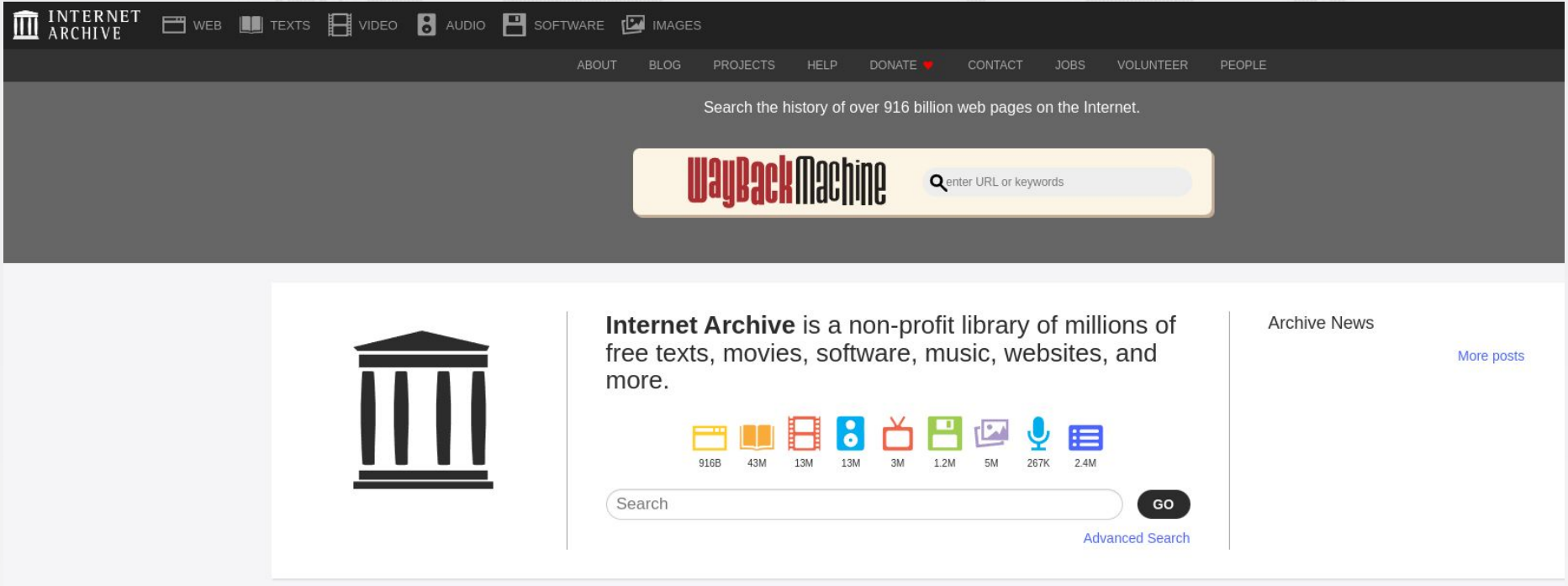


After nine years, Google's open-source code repository, Google Code, started closing shop today by disabling new projects and announcing the permanent shut down of the service by January 26, 2016.

Google Code started as Google's answer to SourceForge, the predominant code repository back in 2006. The reliability of SourceForge was brought into question that year when SourceForge.net's database was hacked and user data was compromised. Problems continued the following summer in 2007 with a temporary service outage in August. The appeal of a repository backed by Google only increased as a result.

ERREUR 404

Archiver le web : la Wayback Machine

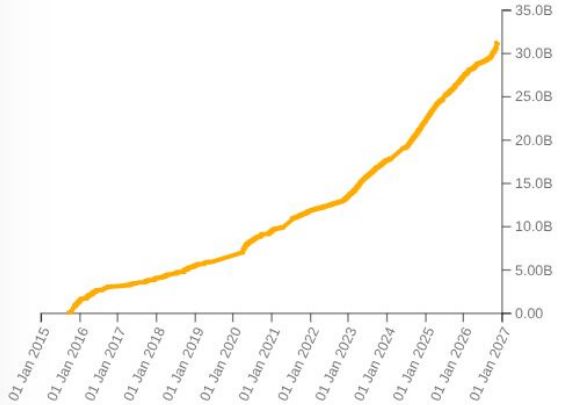


Une archive **universelle** du code source

Collecter, **préserver** et **partager** tout le code source des logiciels

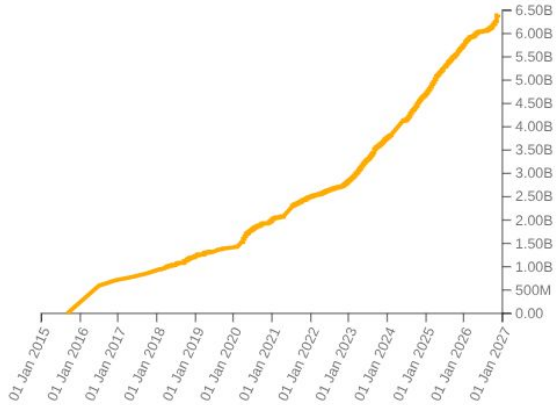
Source files

31,322,182,310



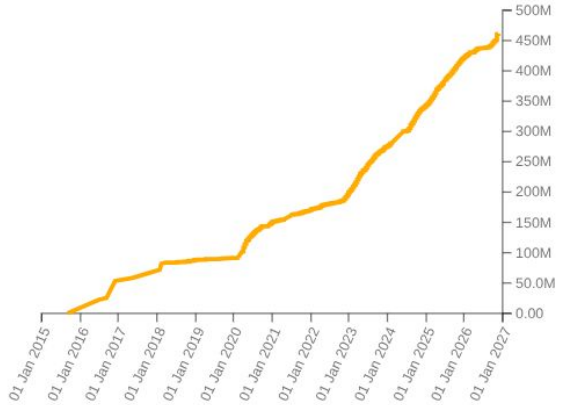
Commits

6,409,466,516



Projects

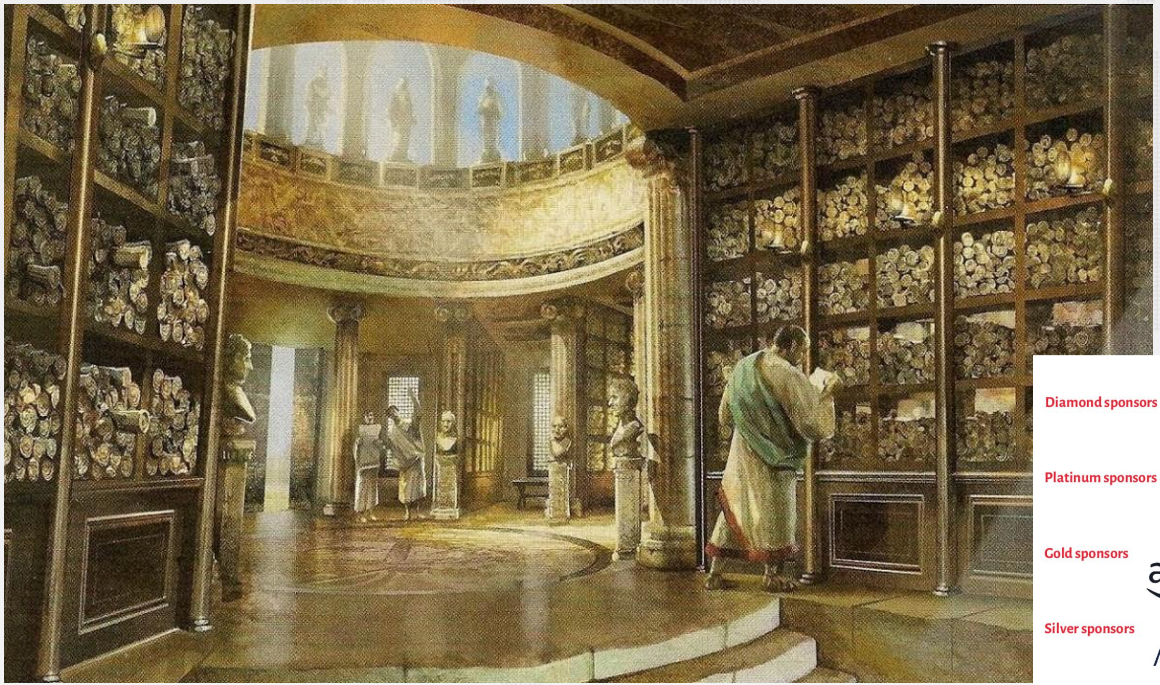
461,282,182



https://archive.softwareheritage.org/

Instantané : 17 nov. 2026

La bibliothèque d'Alexandrie du code source



Lancé en 2016 par

Inria

Fondation SWH hébergée par

Inria FONDATiON

Les mécènes 2025-2026

Diamond sponsors: IBM, Microsoft, TII Technology Innovation Institute

Platinum sponsors: cnrs, Ministère de l'Enseignement supérieur et de la Recherche, Infosys, HUAWEI, CEA

Gold sponsors: aws, Google, openinventionnetwork, Université Paris Cité

Silver sponsors: AdaCore, Premier Ministre, Direction interministérielle du numérique, GitHub, Canadian Research Knowledge Network, Sorbonne Université, Università di Pisa

Bronze sponsors: NWO, Sant'Anna, Scuola Normale Superiore, Université de Lorraine, université Paris-Saclay

En collaboration avec unesco

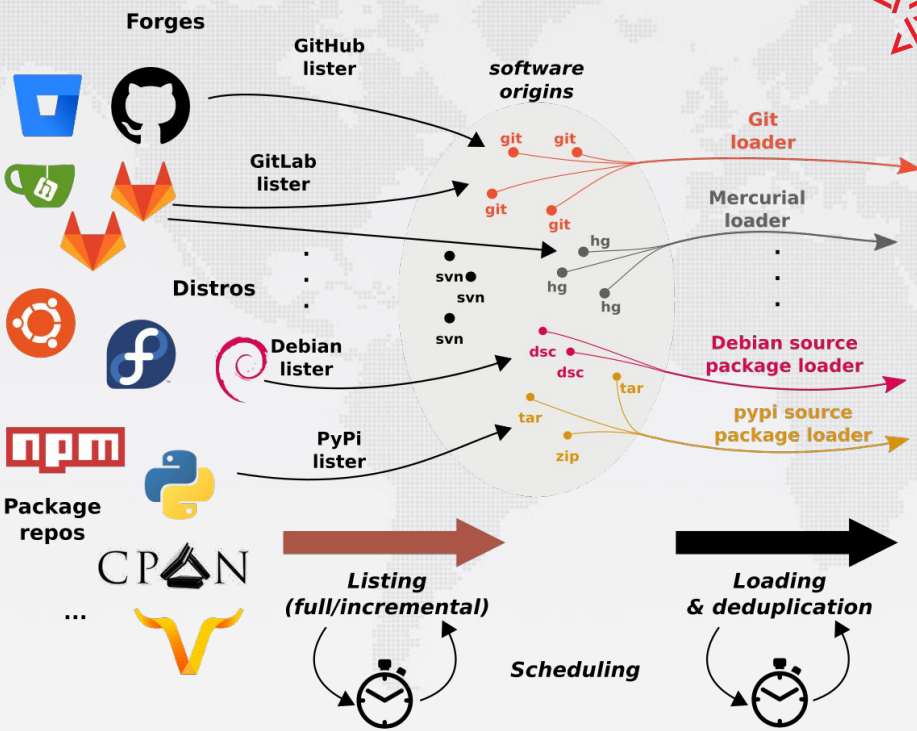
Des forges logicielles à un graphe unifié

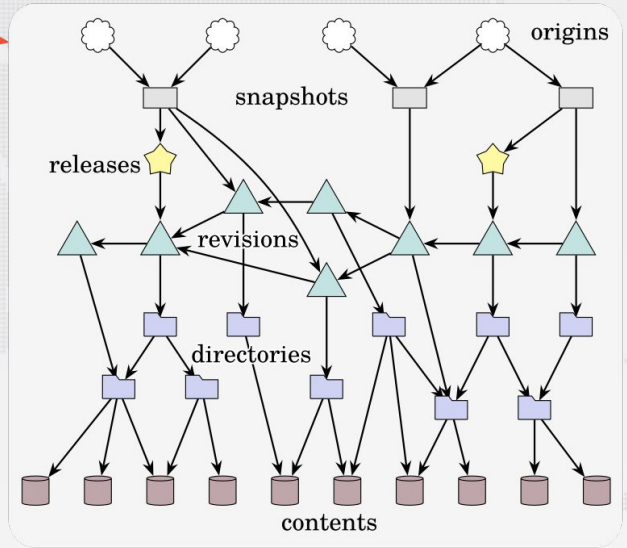
Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Software Heritage Archive

Merkle DAG + blob storage





Le réseau de miroirs de Software Heritage : Une infrastructure distribuée



Construire pour la résilience



Illustration : générée par IA (outils d'images d'OpenAI), novembre 2026.

Contrairement à une idée reçue, la bibliothèque d'Alexandrie n'a pas disparu dans un unique incendie : son déclin fut progressif, au gré des bouleversements politiques, de l'évolution des institutions et de plusieurs destructions au fil des siècles.

« Sauvons ce qui reste : non par des coffres et des verrous qui les soustraient au regard et à l'usage du public en les abandonnant aux ***ravages du temps****, mais par une telle* ***multiplication des copies*** *qu'elle les mette hors de portée de tout* ***accident****. »*

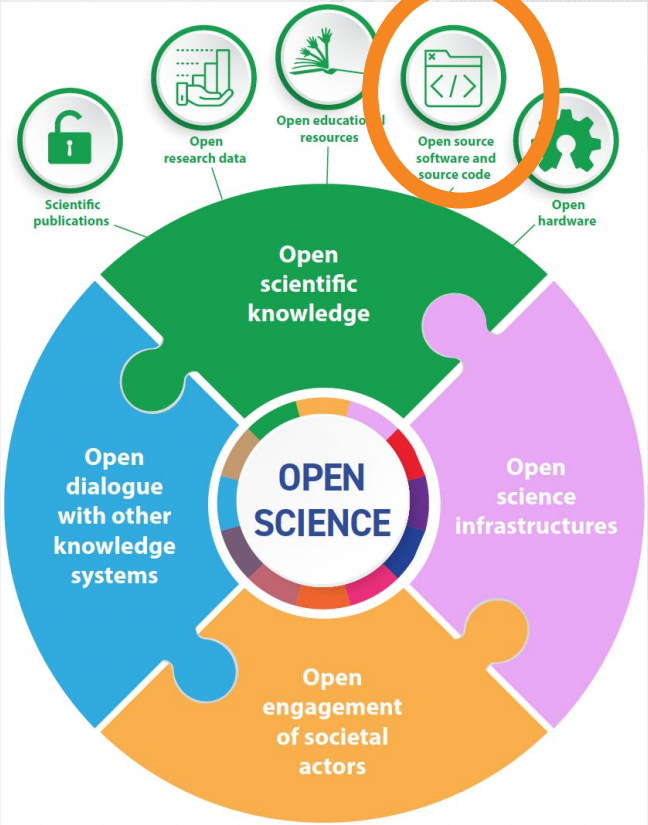
— Thomas Jefferson

Objectifs



- ★ Introduction : le savoir est dans le code source
- ★ Visite guidée : l'archive Software Heritage
- ★ **La science ouverte pour la liberté du savoir**
- ★ IA ouverte, responsable et transparente : un but commun
- ★ Vers la #SouverainetéNumérique

Une vision : la science ouverte pour la liberté du savoir



"S'appuyant sur les principes essentiels de la liberté académique, de l'intégrité de la recherche et de l'excellence scientifique, la science ouverte établit un nouveau paradigme qui intègre dans l'entreprise scientifique des pratiques de **reproductibilité**,
de **transparence**, **de partage** et de **collaboration** résultant de l'ouverture accrue des contenus, des outils et des processus scientifiques."

[Recommandation de l'UNESCO]



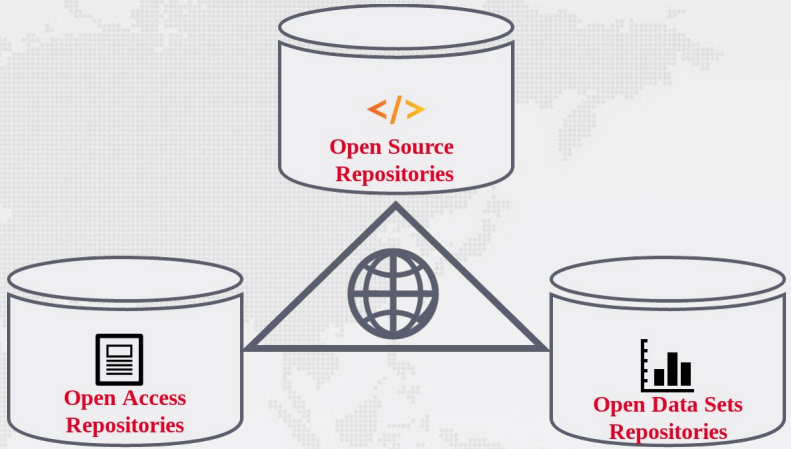
Logiciel de recherche : un pilier de la science ouverte

De multiples facettes, il peut être vu comme :

- un outil
- un produit ou résultat de recherche
- l'objet de la recherche



Tim Berners-Lee : le World Wide Web est né au CERN en 1989, dans la recherche



Les trois piliers de la science ouverte
Gruenpeter, Software Heritage CC-By 4.0 2019

Quels sont les enjeux ?

#DigitalSovereignity

Archiver

➔ pouvoir accéder au logiciel et le récupérer (reproductibilité)

Référencer

➔ pouvoir identifier les artefacts logiciels (reproductibilité)

Citer (pour la reconnaissance)

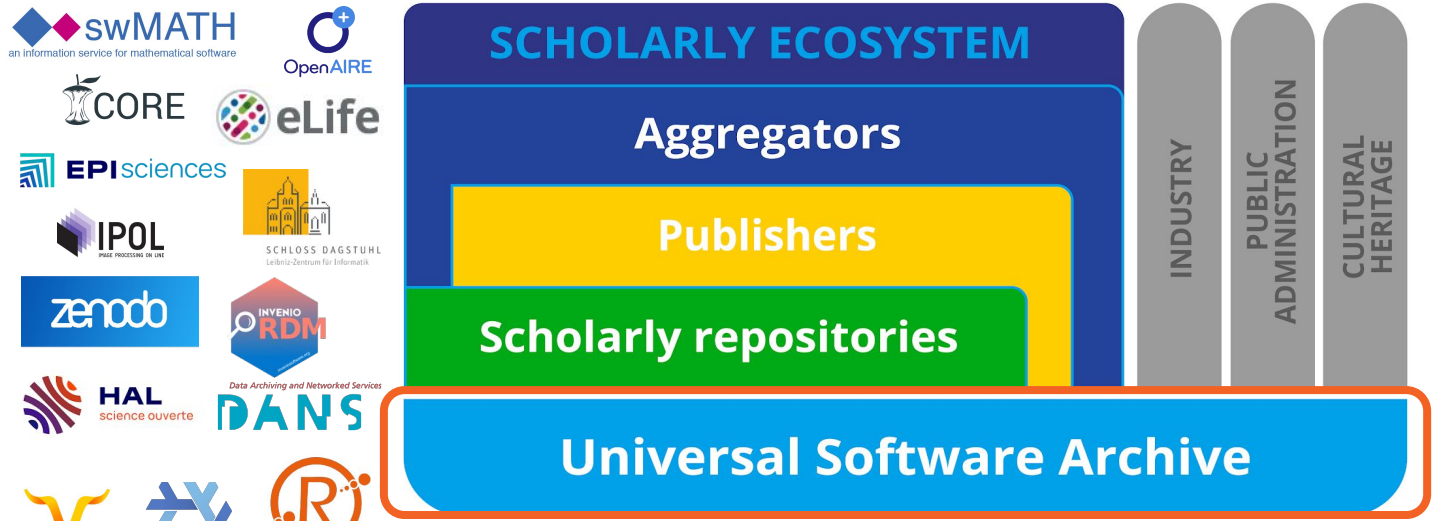
➔ valoriser la création de logiciels en reconnaissant leurs auteurs (évaluation !)

Décrire

➔ faciliter la découverte des projets logiciels (visibilité)

Au service de l'écosystème académique

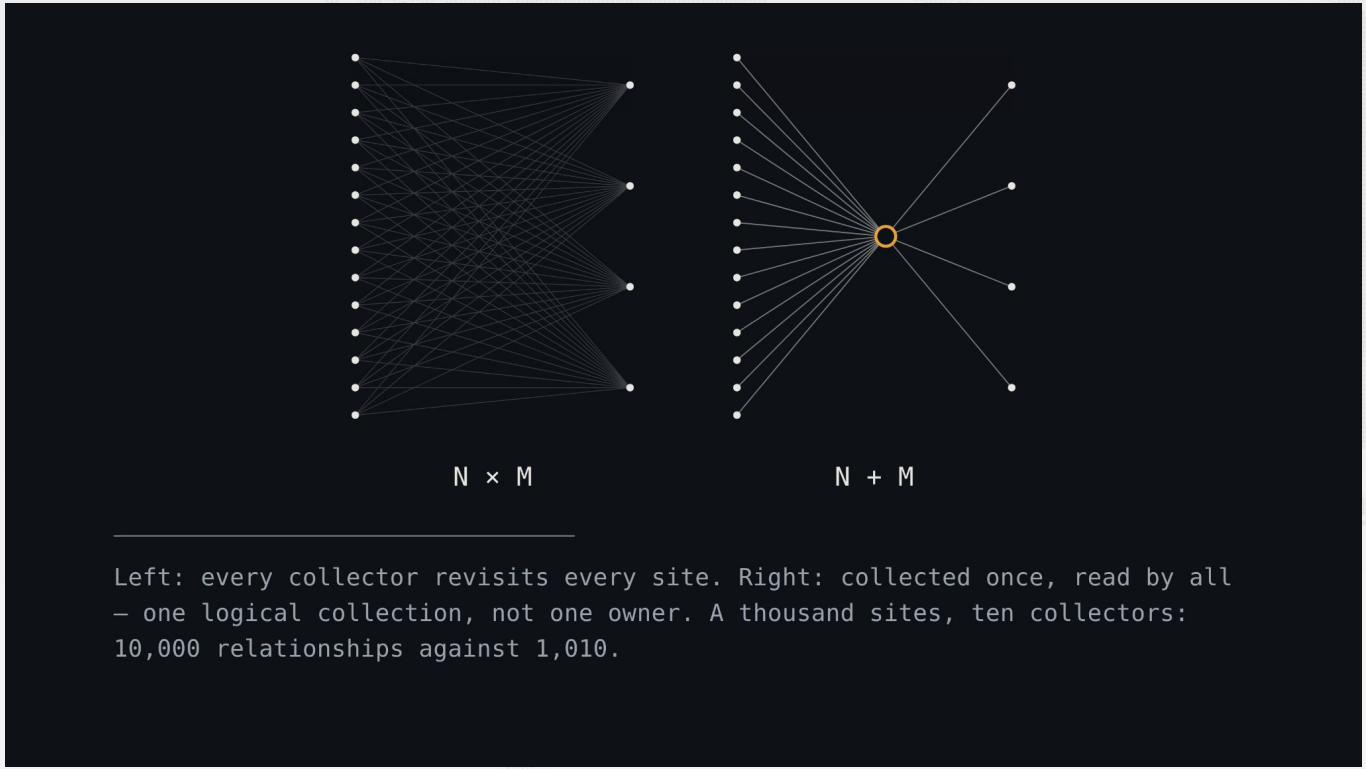
Créer des liens et mutualiser les efforts



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

« Collecter une fois » n'est pas « stocker une fois »



*« Cela ne signifie pas **une seule** machine, **une seule** copie ou **une seule** organisation qui contrôle tout.*

Les données peuvent être répliquées et détenues par des institutions indépendantes relevant de juridictions différentes. »

https://www.dicosmo.org/good-enough/

Objectifs



- ★ Introduction : le savoir est dans le code source
- ★ Visite guidée : l'archive Software Heritage
- ★ La science ouverte pour la liberté du savoir
- ★ **IA ouverte, responsable et transparente : un but commun**
- ★ Vers la #SouverainetéNumérique

Déclaration de Software Heritage sur les LLM pour le code (2023)

- **Ouverture** : Les connaissances issues de l'archive Software Heritage doivent être restituées à l'humanité, et non monopolisées à des fins privées. Les modèles d'apprentissage automatique qui en résultent doivent être mis à disposition sous une licence ouverte appropriée, accompagnés de la documentation et des outils nécessaires à leur utilisation.
- **Transparence** : Les données d'entraînement initiales extraites de l'archive Software Heritage doivent être identifiées de manière complète et précise, par exemple en publiant **les identifiants SWHID** correspondants
- **Mécanisme d'opt-out** : Dans la mesure du possible, des mécanismes doivent être mis en place pour permettre aux auteurs d'exclure leur code archivé des données d'entraînement avant le début de l'entraînement du modèle.

https://doi.org/10.5281/zenodo.24371907

IA en quelques dates



Février 2024: BigCode, The Stack v2 and StarCoder2, Un modèle open-code mis à disposition, entraîné sur le sous-ensemble GitHub de l'archive en **respectant** pleinement les principes.

CodeCommons (**2025-2027**) projet de deux ans financé par l'État français avec des partenaires académiques français et italiens, s'appuie sur Software Heritage pour créer des jeux de données de meilleure qualité au service -d'une IA responsable (en savoir plus).

INFORMATICS EUROPE

ERCIM

Septembre 2026: **Strategic Research and Innovation Agenda (SRIA) for the Future of Software**: « Établir une souveraineté européenne sur les corpus de code utilisés pour entraîner et évaluer l'IA destinée au génie logiciel (par exemple les copilotes et les agents), afin qu'un intrant stratégique de la production logicielle reste sous contrôle européen » (en savoir plus).

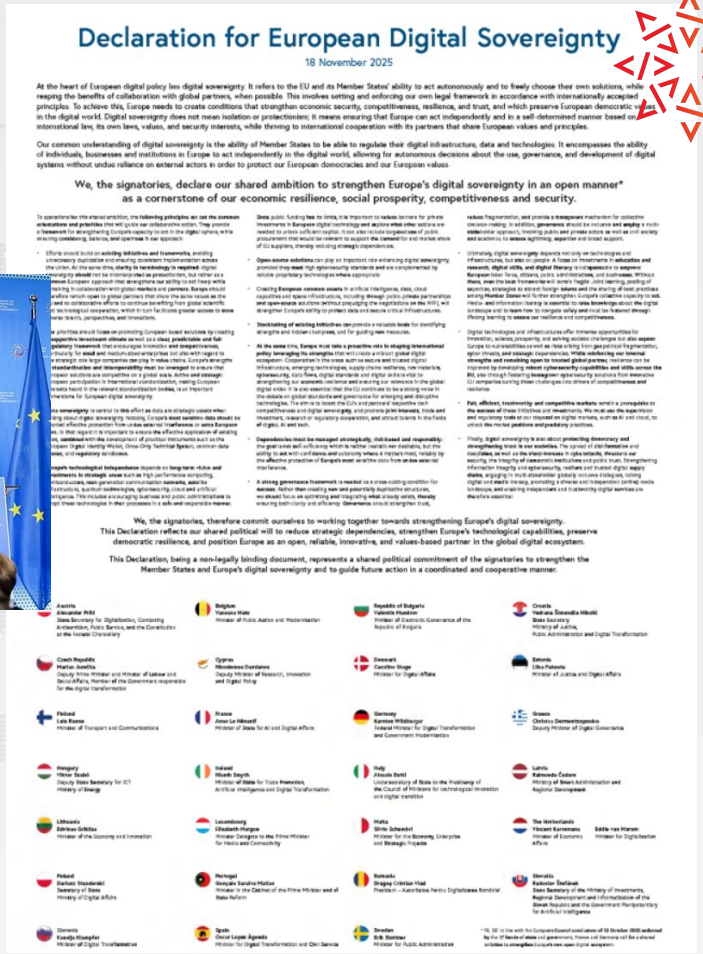
La souveraineté numérique

Points clés de la déclaration de Berlin



- ★ Autonomie stratégique
  - Moins de dépendances
- ★ Communs numériques
  - Adopter l'open source
- ★ **Investissement dans les technologies stratégiques**
  - **IA**, cloud, cybersécurité
- ★ Gouvernance coordonnée et action public-privé

Source : Gouv. fédéral allemand (18/11/25)



Declaration for European Digital Sovereignty

18 November 2025

At the heart of European digital policy lies digital sovereignty. It refers to the EU and its Member States' ability to act autonomously and to freely choose their own solutions, while reaping the benefits of collaboration with global partners, when possible. This involves setting and enforcing our own legal framework in accordance with internationally accepted principles. To achieve this, Europe needs to create conditions that strengthen economic security, competitiveness, resilience, and trust, and which preserve European democratic values in the digital world. Digital sovereignty does not mean isolation or protectionism; it means ensuring that Europe can act independently and in a self-determined manner based on rules, international law, its own laws, values, and security interests, while thriving to international cooperation with its partners that share European values and principles.

Our common understanding of digital sovereignty is the ability of Member States to be able to regulate their digital infrastructure, data and technologies. It encompasses the ability of individuals, businesses and institutions in Europe to act independently in the digital world, allowing for autonomous decisions about the use, governance, and development of digital systems without undue reliance on external actors in order to protect our European democracies and our European values.

We, the signatories, declare our shared ambition to strengthen Europe's digital sovereignty in an open manner* as a cornerstone of our economic resilience, social prosperity, competitiveness and security.

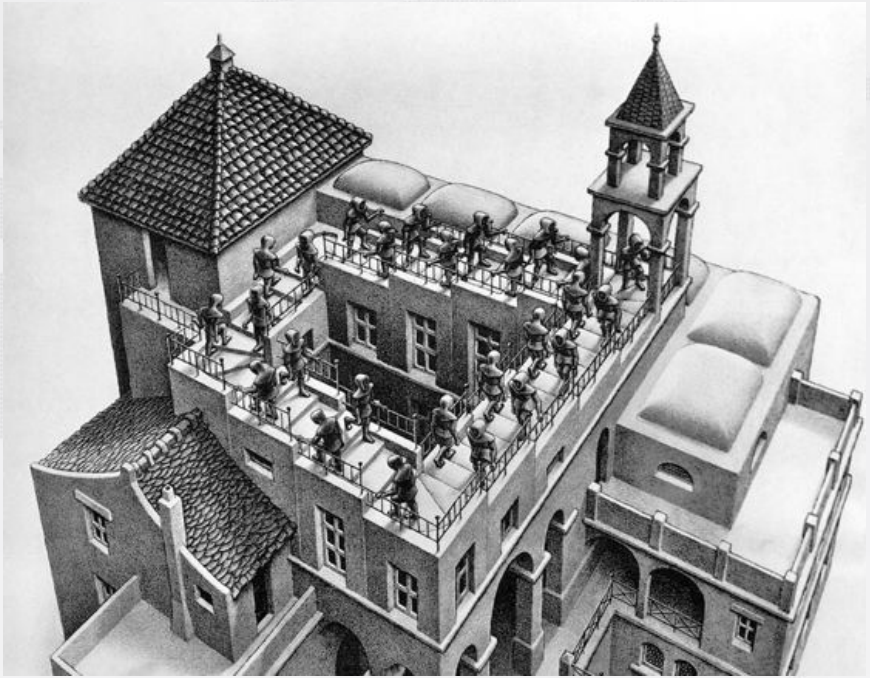
We, the signatories, therefore commit ourselves to working together towards strengthening Europe's digital sovereignty.
This Declaration reflects our shared political will to reduce strategic dependencies, strengthen Europe's technological capabilities, preserve democratic resilience, and position Europe as an open, reliable, innovative, and values-based partner in the global digital ecosystem.

This Declaration, being a non-legally binding document, represents a shared political commitment of the signatories to strengthen the Member States and Europe's digital sovereignty and to guide future action in a coordinated and cooperative manner.

EUtopie ou dystopie pour le logiciel

Comment protéger le savoir ?

Des politiques aux infrastructures :
la boucle étrange est infinie.



Montée et Descente de M. C. Escher

Faire des connaissances scientifiques et techniques représenté dans le code-source un **bien commun public**

Une Europe qui chercherait à *tout posséder* passerait une génération à reconstruire des versions inférieures de choses qui fonctionnent déjà, et n'en serait pas plus libre au bout du compte.

Italo Vignoli

Liberté du logiciel = quand le code source peut être examiné et modifié

La souveraineté n'est pas une question de propriété, mais de liberté.

Objectifs



- ★ Introduction : le savoir est dans le code source
- ★ Visite guidée : l'archive Software Heritage
- ★ La science ouverte pour la liberté du savoir
- ★ IA ouverte, responsable et transparente : un but commun
- ★ **Vers la #SouverainetéNumérique**

La #SouverainetéNumérique naît de la prise de conscience

Reconnaître l'importance du code source : rendre visible l'invisible

- ★ Le code source est un savoir
- ★ Sans le code, la science n'est pas reproductible
- ★ Le logiciel libre et open source n'est pas gratuit
- ★ L'archivage est une solution qui existe déjà

Les institutions doivent **agir** pour soutenir les personnels qui gèrent du code source, et renforcer **la reconnaissance du logiciel de recherche**.

Construire ensemble des écosystèmes interopérables

La reconnaissance du code source exige un effort **soutenu** et **coordonné**. Nous devons traiter le code non comme un simple outil, mais comme un pilier essentiel de notre patrimoine intellectuel.

Mutualisation

SWHID



Reconnaissance

Pérennité

Curation

Ouvert par défaut

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Merci

morane@softwareheritage.org
https://www.softwareheritage.org/

Explorer l'**archive du code source**
https://archive.softwareheritage.org/

Devenir **membre**
https://www.softwareheritage.org/support/members-alig

Devenir **ambassadeur**
https://www.softwareheritage.org/ambassadors/

S'abonner à la **newsletter**
https://www.softwareheritage.org/newsletter/



L'équipe

© Inria / Photo B. Fourrier



La communauté