ÉCONOMIE • HIGH TECH

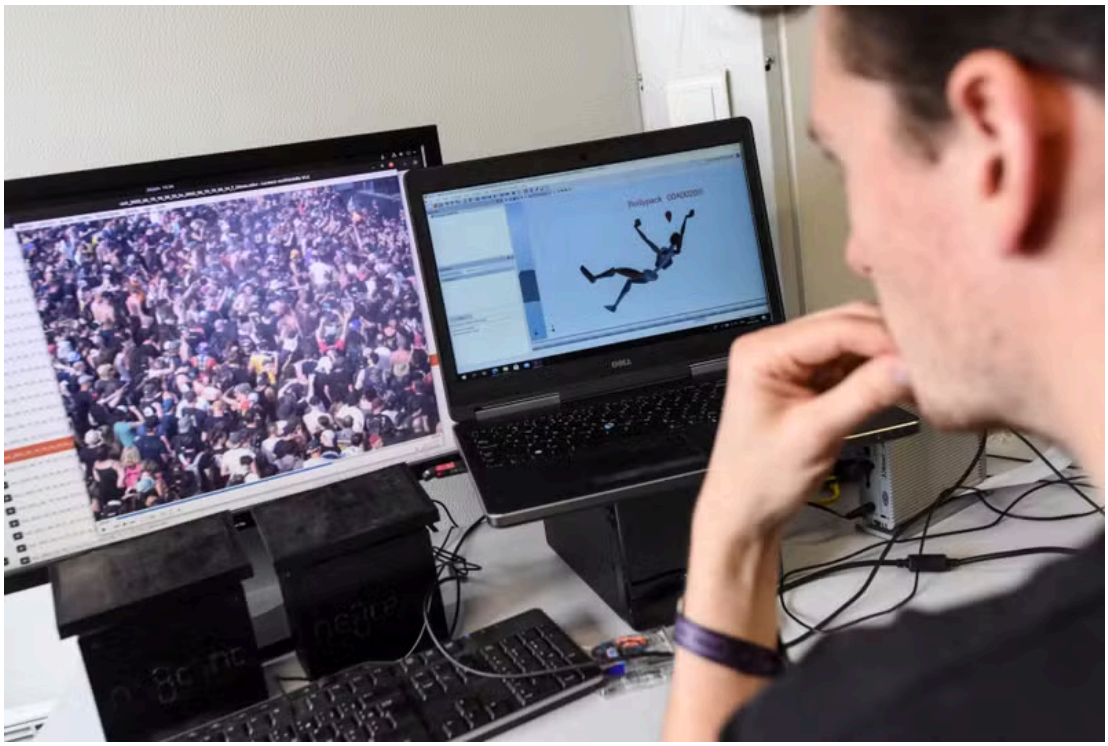
Software Heritage et Scikit-learn, ou la « bibliothèque d'Alexandrie » du code source

Portées par l'Inria, les deux initiatives collectent des milliards de codes au service de la science et de nombreux outils destinés à l'intelligence artificielle.

Par Sophy Caulier

Publié le 05 janvier 2025 à 18h00 · Lecture 1 min.

Article réservé aux abonnés



Etude des mouvements de foule par un scientifique de l'Institut national de recherche en sciences et technologies du numérique étudie, lors d'un concert, à Clisson (Loire-Atlantique), le 26 juin 2022.
SEBASTIEN SALOM-GOMIS/AFP

Les musées conservent les peintures, les bibliothèques les livres, *« mais où sont les codes sources des logiciels sur lesquels désormais tout repose ? »*, interroge Roberto Di Cosmo, professeur d'informatique détaché à l'Institut national de recherche en sciences et technologies du numérique (Inria). Pour répondre à cette question devenue essentielle, Inria a lancé, en 2016, l'initiative Software Heritage, aujourd'hui dirigée par l'informaticien. L'ambition de ce projet d'« archive ouverte » est de *« collecter, préserver et partager tous les logiciels disponibles publiquement sous forme de code source »*, est-il annoncé sur le site Internet du projet.

Lire aussi | L'open source, l'armée de l'ombre du logiciel… et de l'Intelligence artificielle

En moins d'une décennie d'existence, Software Heritage a collecté quelque 22 milliards de codes correspondant à 340 millions de projets. *« Le volume double à peu près tous les deux ans »*, affirme Roberto Di Cosmo. Collectés et vérifiés par des automates, ces codes sont très utiles au monde de la recherche pour la science ouverte, qui consiste à rendre accessibles à tous les données et les résultats des travaux menés.

Ils jouent également un rôle de cybersécurité, car ils constituent une référence standardisée permettant la vérification de l'intégrité des codes, d'en identifier les premiers auteurs, etc. Et, surtout, ils servent à présent à l'entraînement des modèles de langage, si précieux pour les outils d'intelligence artificielle (IA).

Souveraineté numérique

Dans ce domaine, Inria peut se targuer d'un autre succès avec Scikit-learn. Cette bibliothèque regroupe de nombreux outils, algorithmes et fonctionnalités destinés à l'apprentissage automatique (*machine learning*, ML) et accessibles en open source. Depuis son lancement en 2010, Scikit-learn a été téléchargée plus de 1,5 milliard de fois dans le monde, soit plus que les bibliothèques équivalentes de Meta (PyTorch) ou de Google (Tensorflow). Fort de son expertise en la matière, Inria s'est vu confier par le gouvernement français, en mai 2022, la mission de développer *« un ensemble de logiciels ouverts couvrant le cycle de la donnée et des modèles »* et de l'actualiser au fil des innovations.

Lire aussi | De Wikipédia à OpenAI, les communs numériques font de la résistance

Cette mission a donné naissance au projet P16. Partenariat public-privé, ce dernier comporte un volet académique de développement de communs numériques souverains pour l'IA et l'apprentissage automatique. Il comporte également un volet économique et industriel incarné par la start-up Probabl, un éditeur de logiciels open source, qui associe des actionnaires publics, privés et des personnes physiques dans une entreprise à mission de souveraineté industrielle et numérique.

Pour son président-directeur général, Yann Lechelle, l'objectif est *« de produire des dividendes sociétaux en développant les communs numériques open source nécessaires à l'IA et au ML, mais aussi d'atteindre l'équilibre financier en vendant des certifications, du support technique… autour de Scikit-learn »*. Une mission ambitieuse et délicate.

Sophy Caulier