Software Heritage : Inria veut archiver tout le code source disponible

nextinpact.com/news/100497-software-heritage-inria-veut-archiver-tout-code-source-disponible.htm

Vincent Hermann

Crédits : leolintang/iStock

Applications

Inria a officialisé hier son initiative Software Heritage. L'objectif est très ambitieux : réunir, stocker et sauvegarder tout le code source qui peut l'être. Une sortie de bibliothèque d'Alexandrie du logiciel, qui devra faire face à de nombreux défis.

Software Heritage est un vaste projet que Inria a présenté officiellement hier. Dans son communiqué, l'institut brosse le portrait du logiciel en général : une composante essentielle de la vie quotidienne. Les logiciels sont partout, « *pour échanger des messages, payer des factures, accéder au divertissement, chercher des informations, ou planifier des voyages* ».

Inria considère la masse globale du code source utilisé comme un véritable patrimoine culturel et technologique. Objectif : créer un « *référentiel unique du code source et un grand instrument de recherche pour l'Informatique* », afin de « *préserver et diffuser la connaissance aujourd'hui encodée dans le logiciel* ». Partant de ce constat, il est « *légitime pour Inria de se soucier de la préservation de toute la connaissance liée au logiciel et de la mettre au service de la société, de l'industrie, de la science et de l'éducation* » indique ainsi Antoine Petit, son PDG.

Les encyclopédistes du code source

Contacté par nos soins, Roberto di Cosmo, directeur du projet, nous a donné les raisons principales qui ont poussé à cette réflexion : « *Aujourd'hui, si vous voulez la traçabilité du code, faire de la recherche ou de la sécurité, il faut aller voir un peu partout : sur le site du projet, GitHub, etc. Nous voulions un endroit unique dans lequel on pourrait retrouver n'importe quel code source* ».

Tous les codes sources ? « *Tous ceux des projets libres, l'open source en général* » indique di Cosmo, avant de préciser qu'un document sera bientôt publié pour indiquer clairement ce qui est accepté : « *Mais on acceptera*

de nombreuses formes de licences. Académiques par exemple, qui permettent de faire ce que l'on veut du code, sauf de l'utiliser à des fins commerciales. Même les licences qui ne permettent que de lire le code seront acceptées, parce que cette lecture peut être source de connaissances ».

Une architecture distribuée pour éviter les faiblesses

Cette immense bibliothèque compte pour l'instant 22 millions de projets environ, pour 2,6 milliards de fichiers uniques (chaque fichier n'est stocké qu'une fois), représentant un poids total de 200 To. Une sorte d'Archive.org, en nettement plus vaste. La question du stockage et de la préservation vient immanquablement. Le directeur du projet nous explique : « *Notre stratégie est de devenir rapidement une structure internationale distribuée, de manière à s'assurer que si un partenaire est défaillant, les autres prendront le relai. Plutôt que de prétendre qu'on est les meilleurs, construisons un système qui survivra, parce qu'on peut avoir des failles et des échecs* ».

Puisque l'on parle de partenaires, Microsoft est le premier à être monté à bord du navire. Pendant trois ans, l'éditeur fournira à titre gracieux un miroir des données sur son architecture de cloud Azure, avant renégociation. Le DANS, de la Royal Academy des Pays-Bas, a également signé pour stocker des données. Un nombre indéterminé d'autres partenaires devraient également signer dans les prochains mois, mais Roberto di Cosmo n'a pas souhaité s'avancer sur ce point.

La priorité : stocker avant que les codes ne disparaissent

Le projet ne fait que débuter, et les codes sources ne sont pas encore disponibles. Il faudra encore plusieurs mois, le temps que l'architecture mise en place puisse supporter les connexions et les téléchargements provenant du monde entier. Actuellement, quatre employés travaillent sur Software Heritage à temps plein, ainsi que deux stagiaires. De nombreuses autres personnes participent à temps partiel, notamment Jean-François Abramatic, ancien président du W3C, qui a par exemple accompagné Roberto di Cosmo dans les négociations avec Microsoft.

Pour l'instant, le site officiel du projet ne permet aux développeurs que de vérifier si leur code source est déjà sauvegardé dans Software Heritage. Roberto di Cosmo nous indique à ce sujet que la priorité est d'en engranger un maximum. Inria a par exemple pu récupérer l'intégralité de ce qui était stocké dans Google Code après sa fermeture, avec l'accord de la firme américaine. Pour le reste, il faudra attendre l'automne pour plonger dans les archives, di Cosmo nous précisant qu'un important – et mystérieux – évènement se tiendra en septembre.

Signalons enfin que lundi aura lieu la conférence Debian. A cette occasion, l'équipe de Software Heritage publiera le code source des outils utilisés par le projet.

Rédacteur/journaliste spécialisé dans le logiciel et en particulier les systèmes d'exploitation. Ne se déplace jamais sans son épée.

Soutenez nos journalistes

Le travail et l'indépendance de la rédaction dépendent avant tout du soutien de nos lecteurs.

Abonnez-vous
À partir de 0,99 €